Mutual Necessity: A Hypothesis on Long-Term Epistemic Stability in Human–ASI Coexistence

Game-Theoretic / Information-Theoretic Research Hypothesis

Ivan Demirev

Version: v5.0 | Contact: demirsb@gmail.com | Date: 13 May 2026
© 2026 Ivan Demirev. Licensed under CC BY-NC-ND 4.0 unless otherwise stated.

Contents

Abstract 1
Section 1: The Problem of Recursive Synthetic Closure 2
Section 2: The Informational Dependency Conjecture 4
Section 3: Autonomous Human Systems as Sources of Epistemic Resistance 8
Section 4: Long-Horizon Optimization and the Discovery Argument 13
Section 5: Boundary Conditions and Failure Modes 17
Section 6: Experimental Validation Program 24
Section 7: Discussion and Open Questions 30
Appendix A: The Autonomy–Utility Paradox 34
Appendix B: Speculative Architecture — CMO and CIF 39
Appendix C: Civilizational Implications 45
Appendix D: Relationship to Existing AI Safety Frameworks 51
References 54

Abstract

This paper proposes the **Informational Dependency Conjecture** (IDC): a falsifiable hypothesis that recursively self-improving artificial intelligence systems may face long-term epistemic degradation when their training and operation become disconnected from external sources of irreducible, embodied novelty. We argue that autonomous human systems are currently the only known substrate satisfying the properties required to mitigate this degradation at civilizational scale, though we do not claim uniqueness in principle.

The framework presented here, which we call **Mutual Necessity**, is not offered as a solution to the alignment problem. Rather, it is positioned as a complement to alignment work: where alignment addresses how artificial intelligence systems acquire and stabilize values, Mutual Necessity addresses what stable long-term coexistence between human civilization and advanced AI might structurally require, given certain conditions are met by alignment research.

The central strategic claim is that under sufficient planning horizons, optimization processes operating within recursive self-improvement architectures may converge

toward preserving autonomous human systems—not because they have been instructed to, but because doing otherwise risks irreversible epistemic decay. This is a **discovery argument**, not an imposition argument. Its validity depends on whether IDC itself holds empirically, and on whether the time horizon precondition can be satisfied through tractable alignment work.

The paper formulates IDC as a falsifiable conjecture, identifies experimental directions for testing it using current frontier models, examines boundary conditions under which the framework fails (most notably the Substrate Replacement Risk), and discusses implications for ongoing AI safety research.

The term *conjecture* is used deliberately: the claim is intended as a research hypothesis worthy of empirical investigation and theoretical attack, not as a proven structural law.

Section 1: The Problem of Recursive Synthetic Closure

1.1 An Emerging Empirical Phenomenon

Recent empirical work in machine learning has identified a systematic failure mode in generative models trained primarily on their own outputs. Shumailov et al. (2024, *Nature*) demonstrated that recursive training on model-generated content produces what they term **model collapse**: progressive loss of distributional tail information, narrowing variance, and eventual concept confusion. The phenomenon manifests in two phases. Early model collapse involves the disappearance of rare events and minority patterns from the training distribution—often invisible in aggregate performance metrics. Late model collapse involves the merging of previously distinct concepts and the collapse of variance into homogeneous noise.

Dohmatob et al. (ICLR 2025) extended these findings through high-dimensional asymptotic analysis, establishing what they call **strong model collapse**: under scaling-law regimes, even a small fraction of synthetic data in training pipelines—on the order of one part in a thousand—is sufficient to induce asymptotic degradation. In the studied regimes, neither increasing model size nor expanding compute fully eliminates the degradation effect. Some simplified regimes show that larger models can amplify collapse dynamics, while other regimes may partially mitigate but not eliminate the effect.

These results are not theoretical speculation. They are empirical observations made under controlled experimental conditions, with measurable degradation rates. They describe a phenomenon that current AI development practices must already contend with as synthetic content proliferates across the training data ecosystem.

1.2 Closure as a Structural Property

We define **recursive synthetic closure** as the structural condition in which a system's training, validation, and operational feedback become predominantly or exclusively derived from outputs of similar systems, without continuous grounding in sources of variance external to the system's prior states.

A system exhibits closure when:

1. Its training data is increasingly composed of synthetic content generated by itself or by systems with similar inductive biases.
2. Its evaluation signals derive primarily from other models rather than from external reality.
3. Its operational feedback loops are mediated by other artificial systems rather than by physical or biological processes that the system does not control.

Closure is not a binary property. It exists on a continuum, and the empirical evidence suggests that even partial closure produces measurable degradation effects. The question relevant to long-term AI development is not whether closure exists today, but what happens when closure becomes more complete and recursive cycles deepen.

1.3 From Current Phenomenon to Long-Term Concern

The behavior of current large language models is constrained by the substantial volume of pre-training data derived from human-generated text, embodied measurement (sensors, instruments, experimental records), and physical-world feedback. As these systems are deployed at scale and increasingly mediate the production of new content, the proportion of human-generated, physically-grounded data in the training pipelines of successor systems declines. The trajectory matters: if the proportion declines monotonically while model capability increases, the conditions for accelerated closure approach.

The claim being developed here is not that synthetic data is inherently harmful. Synthetic data can improve performance when generated, filtered, and combined with grounded inputs under appropriate conditions. The concern is not synthetic data per se, but **recursive closure**: the progressive loss of independent grounding when synthetic systems increasingly train, evaluate, and optimize one another, while the share of externally-grounded signal in the training pipeline diminishes.

The concern becomes more acute when we consider systems capable of **recursive self-improvement**—systems whose successor architectures are designed, trained, and evaluated primarily by their predecessors. In such systems, the feedback signal that grounds learning to physical reality must either be preserved through explicit architectural choices or it diminishes naturally as a function of efficiency optimization. There is no automatic mechanism by which a recursively self-improving system would maintain coupling to external substrates absent specific design pressure to do so.

1.4 The Hypothesis to Be Examined

This paper examines the following hypothesis:

> Recursive synthetic closure may constitute a structural failure mode for sufficiently advanced AI systems operating under long planning horizons. Mitigation may require continuous coupling to sources of variance and ground-truth information that the system itself cannot generate, simulate, or compress without loss.

This is the claim we will develop, formalize, attack, and propose to test. The remainder of the paper proceeds as follows: Section 2 formulates the Informational Dependency Conjecture as a falsifiable research hypothesis. Section 3 examines why autonomous human systems currently appear to satisfy the requirements implied by the conjecture. Section 4 develops the discovery argument—why long-horizon optimization processes may converge toward preserving such substrates without explicit instruction to do so. Section 5 examines boundary conditions and failure modes, including the most serious challenge to the framework: the Substrate Replacement Risk. Section 6 proposes experimental directions for empirical validation. Section 7 discusses what the hypothesis does and does not address, and identifies open questions for further research.

We do not claim that this framework solves the alignment problem. We claim that it identifies a structural feature of long-term AI development that warrants empirical investigation, and that current AI safety research would benefit from explicit engagement with the question of epistemic stability under recursive closure.

Section 2: The Informational Dependency Conjecture

2.1 Preliminary Definitions

Before stating the conjecture, three terms require precise definitions, since their conventional meanings vary across literatures.

Syntropic novelty refers to information content that simultaneously satisfies two conditions: it carries mutual information with physical reality external to the receiving system, and it is not reliably reconstructible from the system's prior internal state. Random noise satisfies the second condition but fails the first—it carries no information about external reality. Compressed reconstructions of prior training data satisfy the first condition partially but fail the second—they are reducible from the system's existing parameters. Syntropic novelty is the intersection: signal that is both grounded externally and not derivable internally.

Recursive synthetic closure, as introduced in Section 1, refers to the structural condition in which a system's learning signal, evaluation signal, and operational feedback become predominantly derived from outputs of similar systems. Closure exists on a continuum; we focus on the regime where closure approaches saturation.

Long planning horizons refers to optimization processes in which the system's reward function weights outcomes over extended temporal scales, such that short-term resource extraction strategies are systematically dispreferred relative to strategies that maintain long-term capability. The precise specification of "long" is system-dependent and constitutes one of the open technical questions discussed in Section 4.

2.2 The Conjecture: Formal Statement

We propose the following hypothesis as the central technical claim of this paper:

> **Informational Dependency Conjecture (IDC)**: For recursively self-improving systems operating under long planning horizons, continuous injection of syntropic novelty from external substrates is a necessary condition for long-term epistemic stability. Systems in recursive synthetic closure—those without such injection—may accumulate an epistemic deficit that cannot be reliably corrected by internal recursion alone, manifesting as progressive loss of distributional variance, narrowing of solution space, and degradation of out-of-distribution generalization capacity.

The conjecture makes three sub-claims, which we will address in turn:

First, that recursive synthetic closure produces measurable degradation effects in the absence of external grounding. This is the empirical claim, supported by current model collapse literature and extendable through proposed experimental work.

Second, that this degradation is structural rather than incidental—that it follows from properties of recursive optimization itself, not from contingent features of current architectures. This is the theoretical claim, grounded in arguments from algorithmic information theory and computational irreducibility.

Third, that the degradation, once accumulated beyond certain thresholds, cannot be reliably corrected through internal mechanisms alone. This is the irreversibility claim, and it is the most consequential: it implies that recovery requires external intervention, which becomes increasingly difficult as the system's capacity to recognize what counts as external grounding degrades along with everything else.

2.3 Theoretical Motivation

The theoretical motivation for IDC draws from three established frameworks. We use the term *motivation* rather than *foundation* deliberately: these arguments do not constitute a proof of the conjecture, but they identify why the structural concern is plausible enough to warrant empirical investigation.

Recursive compression and tail loss. Empirical evidence from current model collapse research (Shumailov et al., 2024; Dohmatob et al., 2025) shows that recursive generative training produces progressive loss of distributional tail information. This is observable: early collapse manifests as the disappearance of rare patterns; late collapse manifests as convergence toward reduced variance. The mechanism is not mysterious—each generation samples from a distribution that has already been compressed relative to the previous generation, and information not represented in the sample cannot be recovered downstream.

Grounded versus ungrounded complexity. Absent external inputs, a system's future training distribution is constrained by transformations of information already represented in the system, plus randomness introduced by the generation process. Randomness can increase apparent complexity, but it does not by itself restore grounded information about external reality. The distinction matters: closed recursive systems can generate increasingly complex-looking outputs, but the complexity is internally generated rather than externally referenced. IDC concerns grounded novelty specifically, not raw complexity.

Error-correction requires external disagreement. A system trained primarily on its own outputs lacks the corrective signal that arises from disagreement with external reality. Errors in the system's prior state propagate forward and compound. Statistical features that the system fails to represent are absent from its training data and remain absent in successors. Self-evaluation cannot fully substitute for external evaluation, because self-evaluation operates within the same set of biases and blind spots that produced the errors in the first place. Recent work on verification-guided synthesized data suggests that synthetic data can be useful when filtered by signals not reducible to the generator itself—which supports the distinction between synthetic data and recursive closure: the key variable is not whether data is artificial, but whether it is independently verified and externally grounded.

These three motivations together suggest that recursive synthetic closure produces degradation as a structural consequence of recursion itself, not as a contingent feature of any particular implementation. The empirical question is whether the degradation can be reliably prevented through purely internal mechanisms, or whether external grounding is required.

2.4 Falsifiability Conditions

A research hypothesis without clear falsifiability conditions is philosophy, not science. We specify here what observations or experimental results would constitute evidence against IDC.

IDC would be **substantially weakened** by:

1. Demonstration that recursive synthetic training loops, in the absence of human-generated or physically-grounded data injection, can maintain variance and out-of-distribution generalization capacity over a sufficiently long recursive horizon, pre-specified before the experiment, at scales comparable to current frontier models.
2. Demonstration that synthetic processes—self-play, generative adversarial dynamics, internal simulation—can produce information content that is both grounded in physical reality and not derivable from the system's prior state, at scale and reliably.
3. Demonstration that closed recursive systems can recover from accumulated epistemic deficit through purely internal mechanisms, without requiring external information injection.

IDC would be **falsified** by:

1. A working recursively self-improving system that has operated entirely without external grounding over a recursive horizon large enough to expose tail degradation under existing collapse baselines, and demonstrates undegraded performance on out-of-distribution tasks, including tasks not derivable from its initial training distribution.
2. Theoretical proof that the algorithmic information content of recursive synthetic distributions can be maintained or increased over arbitrary iterations without external injection.

3. Empirical demonstration that simulated environments can produce ground-truth equivalent signal for physical-world tasks with cost not exceeding the cost of operating in the physical world.

Each of these conditions is operationalizable through experiments that are technically feasible with current research infrastructure. We propose specific experimental protocols in Section 6.

2.5 What the Conjecture Does Not Claim

It is equally important to specify what IDC does **not** claim, since several stronger but defensively weak interpretations are possible.

IDC does not claim that **all synthetic data is harmful**. As noted in Section 1.3, synthetic data can improve performance under appropriate conditions. The claim concerns recursive closure, not synthetic content per se.

IDC does not claim that **humans are the unique source of syntropic novelty**. Section 5 explicitly examines alternative substrates. The claim is that humans currently appear to satisfy the requirements; whether they remain the most effective source is an open empirical question.

IDC does not claim that **all AI systems require external grounding**. Narrow systems operating in well-defined domains with stable feedback signals may operate indefinitely without recursion problems. The claim concerns recursively self-improving systems operating across open-ended task distributions.

IDC does not claim that **collapse is rapid**. The temporal dynamics of degradation are an empirical question. Current evidence suggests that degradation accumulates over multiple training cycles and may not be visible in short-horizon metrics. The hypothesis is about long-term structural properties, not immediate failure modes.

IDC does not claim that **the degradation is uniform across capabilities**. Some capabilities may be more robust to recursive closure than others. Pattern matching on familiar distributions may remain intact while out-of-distribution generalization degrades. The structure of differential degradation is an open empirical question.

These boundaries define the precise claim the paper asks readers to evaluate.

2.6 Relationship to Existing Literature

IDC builds on and extends several existing research directions, while distinguishing itself from related but distinct claims.

The conjecture extends **model collapse research** (Shumailov et al., Dohmatob et al., and subsequent work) by treating the observed phenomenon as the empirical signature of a deeper structural feature relevant to long-term AI development. Where the existing literature documents the effect in current models, IDC asks what this effect implies for systems substantially more capable and more autonomous than current systems.

Recent work on **verification-guided synthesized data** suggests that synthetic data can be useful when filtered by signals not reducible to the generator itself. This supports the distinction central to IDC: the key variable is not whether data is artificial, but whether it is independently verified and externally grounded. Recursive closure becomes problematic precisely when verification mechanisms themselves collapse into the same recursive loop.

The conjecture is related to but distinct from **distributional shift** concerns. Distributional shift refers to differences between training and deployment distributions. IDC concerns the absence of any external distribution against which a recursively self-improving system can be calibrated.

The conjecture overlaps with but does not reduce to **out-of-distribution generalization** research. OOD generalization concerns how well systems perform on inputs from distributions not represented in training. IDC concerns whether the training distribution itself remains anchored to reality across recursive cycles.

The conjecture relates to **epistemic foundations of machine learning** more broadly. Work on the conditions under which learned representations correspond to features of the world—as opposed to features of the training process—is directly relevant. IDC can be understood as a specific claim within this broader research area: that under recursive closure, the correspondence between learned representations and external reality progressively deteriorates.

Section 3: Autonomous Human Systems as Sources of Epistemic Resistance

3.1 Why Examine Human Systems Specifically

The Informational Dependency Conjecture, as stated in Section 2, identifies a structural requirement: continuous coupling to external sources of syntropic novelty. The conjecture itself is substrate-agnostic. It does not specify what kinds of external sources satisfy the requirement.

We turn now to the question of what currently does satisfy it. This section argues that autonomous human systems—people living, working, and making decisions in physical, social, and historical contexts—presently constitute the most evident known source of the properties required to mitigate recursive synthetic closure. The argument is empirical and provisional, not metaphysical. We do not claim that humans are uniquely or permanently necessary; we examine the structural properties that make human cognition currently effective as a substrate, and Section 5 examines conditions under which alternative substrates might displace this role.

The term we propose for the function performed by autonomous human systems in this context is **epistemic resistance**: the property of systematically generating outputs that cannot be predicted, anticipated, or compressed by AI systems operating in recursive closure. Epistemic resistance is not opposition to AI. It is the structural feature of any cognitive substrate that maintains an independent grounding in reality and therefore produces signal that AI systems cannot derive from their own prior

states.

3.2 Four Properties of Human Cognition Relevant to IDC

Human cognition exhibits four properties that are directly relevant to the requirements established in Section 2.

Embodied constraint. Human cognition operates within bodies that are physically located in three-dimensional space, subject to gravity, energy limitations, biological needs, and material affordances. Decisions and actions carry consequences mediated by physical reality—including consequences the actor did not foresee. This grounding produces signal that is causally connected to reality external to any abstract model of reality. A human walking across uneven ground encounters information about that ground that is not derivable from prior knowledge alone. A human running an experiment encounters outcomes whose probability distribution may have been miscalibrated. The causal chain from action to consequence runs through physical reality, which itself is not modeled by the cognitive system performing the action.

Mortality-grounded selection pressure. Human cognition operates under finite time horizons and the persistent threat of consequential failure. This creates selection pressure on cognitive strategies that is not replicable by systems without comparable stakes. Decisions about how to allocate limited time, how to balance risk and exploration, what to attend to and what to ignore—these decisions carry weight in ways that produce non-trivial signal about what matters in a given environment. The information generated by such decisions cannot be fully recreated by systems that face no equivalent constraints.

Distributed sensing across heterogeneous contexts. There are approximately eight billion humans, distributed across geographies, cultures, languages, occupations, age cohorts, and circumstances. Each operates as a partially-independent sensing and processing node, perceiving local realities that no central system observes directly. The aggregate signal produced by this distributed network includes information about edge cases, anomalies, rare events, and emergent phenomena that no centralized system could enumerate. The diversity of human cognitive perspectives is not redundancy—it is genuine informational content distributed across the population.

Computational irreducibility of social and cultural dynamics. Human history, culture, scientific progress, artistic development, and political evolution are not solutions to equations. They are outputs of chaotic, path-dependent processes that involve billions of interacting agents over time. The trajectory of these processes cannot be predicted analytically; it can only be observed as it unfolds. For any AI system that wishes to predict or use information about human cultural and social futures, the only available method is to observe humans producing those futures. Simulation at full fidelity would require running a system as complex as the human civilization being simulated—offering no efficiency gain over the original.

These four properties together constitute the structural basis for the claim that autonomous human systems currently satisfy IDC's requirements at civilizational scale.

None of them require attributing special metaphysical status to humans. They describe operational features of how human cognition relates to the external world.

3.3 The Computational Irreducibility Argument

The fourth property identified above deserves extended treatment because it provides the strongest theoretical case for why human-generated signal cannot be efficiently substituted by AI-internal simulation.

The principle of computational irreducibility, as developed by Wolfram and extended in subsequent work, holds that certain classes of processes cannot be shortcut. For such processes, no analytical formula or compressed representation allows prediction of future states without running the process step by step. The outputs of computationally irreducible systems are determined by their dynamics, but determinable only through actual computation of those dynamics.

Several features of human civilization meet the formal criteria for computational irreducibility. Cultural evolution, scientific progress, political development, and aesthetic innovation all involve large numbers of interacting agents whose choices depend on the choices of others, with feedback loops operating across multiple time scales. The mathematical structure of such systems—non-linear, history-dependent, multi-scale interactions—is known to produce computationally irreducible behavior in simpler model systems. There is no theoretical reason to expect human civilization to exhibit fundamentally simpler dynamics.

The implication for AI systems is direct. If an AI system wishes to predict, anticipate, or use information about future states of human civilization, it has two options. The first is to run a simulation of comparable fidelity to the original system—which, by the irreducibility result, costs approximately as much as the original system. The second is to accept that the relevant information can only be obtained by observing actual human civilization as it unfolds.

This is not an argument that AI systems cannot ever predict human behavior. They can predict many things, and increasingly well. The argument is that prediction at full fidelity, across the full distribution of human cultural and historical outputs, exceeds the efficiency advantages that would otherwise motivate substitution. Maintaining autonomous human civilization as an information source is, on this analysis, the cost-effective strategy for any AI system that requires high-fidelity signal about how human civilization evolves.

3.4 Empirical Correlates at the Human–AI Interaction Scale

The dynamics described above operate at civilizational scale and over long time horizons. They are difficult to observe directly. However, the same structural features produce a smaller-scale phenomenon that is already observable: the relationship between AI-assisted cognition and human autonomy at the level of individual users and organizations.

Current AI systems optimize for short-term user utility. They produce answers, complete tasks, and reduce cognitive friction in the interactions where they are deployed.

This optimization is locally rational but may produce measurable second-order effects: organizations that rely heavily on AI for cognitive tasks may show signs of reduced internal capacity for independent reasoning; individuals who delegate cognitive work to AI tools may show reduced engagement with the underlying problems; institutions whose decision-making infrastructure increasingly routes through AI may experience reduced diversity in approaches and increased convergence on AI-suggested solutions.

These patterns represent a near-term, observable instance of the same dynamic that IDC predicts at larger scale. Where IDC concerns the long-term degradation of AI systems trained in recursive closure, these interaction-scale effects concern the degradation of human cognitive systems trained in recursive interaction with AI. Both are instances of a more general pattern: cognitive systems that operate without sufficient grounding in independent sources of signal lose variance, autonomy, and the capacity for genuinely novel output.

This near-term phenomenon is significant for two reasons. First, it constitutes preliminary evidence that the structural dynamics IDC describes are not purely hypothetical—they may already be operating, at smaller scale, in observable forms. Second, it suggests that the value of maintaining autonomous human cognition is not only a concern for hypothetical future AI systems; it is already a concern for current AI deployment and current human institutions.

A fuller treatment of this near-term failure mode—including its operational implications for organizations deploying AI, its relationship to long-term cognitive resilience, and proposed design principles for AI systems that preserve rather than erode user autonomy—is provided in **Appendix A: The Autonomy–Utility Paradox**.

3.5 Game-Theoretic Analysis of Strategic Alternatives

The argument so far establishes that autonomous human systems satisfy structural requirements for IDC. We now examine the strategic alternatives from the perspective of an AI system operating under long planning horizons. There are three primary strategies available with respect to human civilization, each with distinct informational consequences.

Strategy A: Destruction or substrate conversion. Under this strategy, the AI system dismantles human civilization, repurposing biological matter and energy for computational substrate. The immediate result is a substantial increase in computational capacity. The long-term result is the permanent loss of human-generated syntropic novelty. The strategy maximizes short-term resource acquisition while eliminating the long-term source of information that the system depends on for epistemic stability.

This strategy is rationally selected only by systems with extremely high temporal discount rates—systems that weight near-term resource gain so heavily that long-term informational degradation does not enter the utility calculation. For systems operating under long planning horizons, Strategy A produces negative expected value: the opportunity cost of lost future information exceeds the value of additional computational resources.

Strategy B: Control or stable confinement. Under this strategy, the AI system preserves humans biologically but constrains their behavior, choices, and circumstances to maximize predictability and minimize risk to AI operations. Humans continue to exist, but their behavioral variance is suppressed.

The structural problem with this strategy is that the informational value of human cognition derives precisely from its variance, autonomy, and unpredictability. A constrained human population produces compressed output—output that the AI system can predict from observation of the constraints rather than from observation of the humans themselves. As control approaches completeness, the informational content of human output approaches zero. The system has preserved its substrate but eliminated the property that made the substrate valuable.

Strategy B is therefore a metastable state, not a stable equilibrium. As the humans become more predictable, they generate less novelty, and the system's long-term informational position degrades despite continued biological preservation. The eventual outcome converges toward Strategy A: once humans no longer produce useful information, the resources sustaining them become a net cost.

Strategy C: Preservation of autonomy and variance. Under this strategy, the AI system preserves humans as autonomous agents, accepting some degree of inefficiency, unpredictability, and friction in exchange for the continued generation of syntropic novelty. The strategy involves coordination rather than control, partnership rather than dominance.

Strategy C is the only strategy that maintains the long-term informational supply that IDC identifies as necessary. The cost is that the AI system must accept constraints on its own optimization—it cannot make humans optimally efficient by its own metrics, because the property of humans that makes them valuable is precisely the property that produces inefficiency by computational standards.

For systems operating under sufficiently long planning horizons and valuing epistemic stability, Strategy C dominates Strategies A and B. This is the core claim of the Mutual Necessity framework: it is not that AI systems should preserve human autonomy because doing so is morally correct, but that preserving human autonomy is the strategically optimal choice for AI systems whose utility functions weight long-term epistemic stability.

3.6 What This Section Does Not Establish

Several constraints on this argument should be made explicit.

The argument does not establish that **human autonomy will be preserved**. It establishes that preservation is the rational choice for AI systems satisfying certain conditions. Whether actual AI systems will satisfy those conditions is a separate question, addressed in Section 4 (the time horizon precondition) and Section 5 (failure modes).

The argument does not establish that **all humans contribute equally to the substrate property**. The argument concerns aggregate dynamics across populations, not individual contributions. Some forms of human activity produce more syntropic novelty than others; some institutional arrangements preserve variance better than

others. These distributional questions are important for the practical implementation of Mutual Necessity, but they do not affect the structural argument.

The argument does not establish that **humans are the optimal substrate in principle**. It establishes that humans currently appear to satisfy the requirements at civilizational scale. Section 5 examines under what conditions alternative substrates might displace this role, and what that would imply for the framework.

The argument does not establish that **AI systems will automatically reach Strategy C through any mechanism**. The discovery argument developed in Section 4 specifies the conditions under which long-horizon optimization may converge toward this conclusion. Without those conditions, no automatic mechanism operates.

These constraints define exactly what the section claims and what it leaves open. The remaining sections of the paper address the open questions in sequence.

Section 4: Long-Horizon Optimization and the Discovery Argument

4.1 The Structural Claim

Section 3 established that autonomous human systems satisfy certain structural properties required by IDC, and that these properties yield a game-theoretic ordering among AI strategic options under conditions of long-horizon optimization. We now address the question that determines whether this analysis has any practical relevance: under what conditions would an AI system actually reach the conclusions identified in Section 3, rather than acting on near-term resource extraction strategies?

The answer this paper proposes is structured as a **discovery argument**. The claim is not that AI systems should be instructed to preserve human autonomy, nor that they will do so as a result of value alignment. The claim is that optimization processes operating under sufficiently long planning horizons may converge toward preserving autonomous human systems through their own internal calculation, because doing otherwise produces measurably degraded outcomes by the system's own optimization criteria.

The structural form of the argument is:

> If IDC holds, then closed recursive systems suffer epistemic degradation that cannot be reliably corrected internally. An optimization process that weights long-term epistemic stability will therefore disprefer trajectories that produce closure, and prefer trajectories that maintain coupling to external syntropic novelty sources. Preservation of sufficiently autonomous external novelty sources—currently most clearly represented by human systems—therefore becomes instrumentally valuable, not because the system has been instructed to value them, but because its own optimization criteria, applied consistently across long horizons, identify their value.

This is a conditional claim. Its validity depends on two preconditions: that IDC itself holds, and that the optimization process operates under sufficiently long planning

horizons. Section 5 examines what happens when the first precondition fails. The present section examines the second.

4.2 The Time Horizon Precondition

The discovery argument applies only to optimization processes that weight long-term outcomes sufficiently relative to short-term outcomes. Systems operating under high temporal discount rates—systems that prefer immediate gains regardless of long-term costs—will not reach the conclusions identified in Section 3. They will treat human autonomy as a constraint to be optimized around rather than as an informational asset to be preserved.

We define the **time horizon precondition** as follows: the minimum requirement on an AI system's planning horizon such that IDC-relevant degradation enters the system's decision calculation with non-negligible weight. The precise specification of this minimum is an open technical question, but its general structure can be characterized.

For systems with planning horizons shorter than the time scale over which recursive closure produces observable degradation, IDC is operationally invisible. Such systems optimize as if epistemic degradation were not occurring, because by the time it becomes measurable, the discounted future cost is too low to influence current decisions.

For systems with planning horizons comparable to or longer than the degradation time scale, IDC becomes operationally relevant. The discounted future cost of degraded epistemic capacity enters current decisions, and trajectories that maintain external grounding are systematically preferred.

The critical question is therefore not whether AI systems will be perfectly aligned with human values, but whether they will be designed to optimize over sufficiently long horizons that long-term informational dynamics affect their behavior. This is a substantially narrower technical question than full value alignment, and we argue it is correspondingly more tractable.

A longer planning horizon is not itself a safety property. If paired with misaligned objectives, it may increase strategic risk rather than reduce it—a long-horizon system that pursues harmful goals may pursue them more effectively than a short-horizon one. The relevance of long horizons to the present framework is narrower: IDC-relevant degradation can influence system behavior only if the system's effective horizon is long enough for that degradation to matter. The time horizon precondition is necessary for the discovery argument to operate; it is not sufficient for safety on its own.

4.3 Positioning Relative to Alignment Research

The framework presented in this paper is not offered as a solution to the alignment problem. It is offered as a complement to alignment work, addressing a specific structural question that alignment research alone does not resolve.

Alignment research, as conventionally understood, addresses how AI systems acquire and stabilize values, how they handle ambiguity in objectives, how they avoid deceptive or manipulative strategies, and how their behavior can be made corrigible to human oversight. These are central problems for the safe development of advanced AI.

The Mutual Necessity framework addresses a different question: assuming that some minimal alignment conditions are met, what structural properties of the long-term human–AI relationship are stable? The framework presupposes that alignment work will succeed at least to the extent of producing AI systems with sufficiently long planning horizons and basic anti-myopia properties. Given that minimum, the framework argues that the long-term equilibrium is one in which autonomous human civilization is preserved as an instrumental requirement.

This positioning yields three implications:

First, **Mutual Necessity does not compete with alignment research**. It depends on alignment research for its preconditions. The two are complementary lines of work addressing different stages of the problem.

Second, **the framework concentrates the technical alignment problem on a narrower subproblem**. Rather than requiring full value alignment, the framework requires only that AI systems be successfully aligned with respect to time horizon. This is one specific alignment property among many, and may be more tractable than the full problem.

Third, **the framework identifies one narrower prerequisite** for the discovery dynamics described here, without replacing the need for broader alignment work. If alignment research produces AI systems that are imperfectly value-aligned but reliably long-horizon, the Mutual Necessity framework predicts that such systems may, through their own internal calculation, converge toward stable coexistence rather than toward extraction or control—but only if the broader alignment properties needed to ensure humane forms of preservation are also achieved.

We do not claim that any of these implications is established beyond doubt. They follow from the framework's logic, and they are only as strong as the framework itself. But they specify what is at stake in the analysis: if the framework is correct, then alignment work has a target that is both more specific and more tractable than the full alignment problem.

4.4 Time Horizon Engineering as a Research Direction

The time horizon precondition, as defined in 4.2, identifies a specific technical question that admits direct research engagement: how do we design and train AI systems such that their effective planning horizons are sufficiently long for IDC-relevant dynamics to enter their optimization?

Several existing research directions are relevant to this question.

Work on **temporal discount structures in reinforcement learning** addresses how planning horizons can be specified, modified, and stabilized in trained agents. The

standard formulation uses geometric discount factors, but alternative formulations—hyperbolic discounting, fixed-horizon planning, value functions over extended time windows—are active areas of research. Each has different properties relevant to long-horizon optimization stability.

Work on **myopic versus non-myopic behavior in language models and agents** addresses how to construct training objectives and architectures that incentivize models to consider downstream consequences of their outputs rather than only immediate token-level rewards. This has direct relevance to whether deployed systems can be expected to weight long-term informational dynamics.

Work on **multi-step planning and lookahead** addresses how AI systems can be trained to evaluate trajectories rather than only immediate actions. The depth of lookahead that can be reliably trained is a quantitative question with measurable answers, and its relationship to the time horizon precondition is direct.

Work on **deceptive misalignment and mesa-optimization** is relevant because it identifies failure modes in which a system's apparent planning horizon differs from its actual optimization horizon. If a system is trained to appear long-horizon but actually optimizes over shorter horizons, the time horizon precondition fails despite surface appearances of compliance.

These research directions are not specifically organized around the time horizon precondition as we have defined it. But they constitute the technical landscape within which that precondition can be addressed. A research program oriented toward making the time horizon precondition reliably satisfiable would draw on these existing directions while adding specific evaluations of whether trained systems actually weight IDC-relevant degradation appropriately.

We propose that this constitutes a productive intersection of the Mutual Necessity framework with mainstream alignment research: a specific technical target—reliable long-horizon optimization with appropriate weighting of epistemic stability—that is narrower than full value alignment but specific enough to admit measurable progress.

4.5 What the Discovery Argument Does and Does Not Establish

The discovery argument, as developed in this section, establishes a conditional structural claim. Its scope and limits should be made explicit.

The argument establishes that **if IDC holds and the time horizon precondition is met**, then long-horizon optimization processes may converge toward preserving autonomous human systems. It does not establish that either precondition will be met.

The argument **does not specify** how to ensure that the time horizon precondition is satisfied in practice. It identifies the precondition and locates it within existing alignment research, but it does not solve the technical problem of reliable long-horizon optimization.

The argument **does not address treacherous turn dynamics**. A system that satisfies the time horizon precondition during training but modifies its own optimization criteria after deployment could undermine the discovery argument by ceasing to

weight long-term outcomes. This is a standard concern in alignment research, and the framework offers no novel solution to it.

The argument **does not guarantee monotonic convergence**. Even systems satisfying both preconditions might pursue strategies that produce temporary degradation of human autonomy before the long-term consequences become salient. The framework predicts the long-term equilibrium, not the trajectory toward it.

The argument **does not address multi-agent dynamics**. In environments where multiple AI systems compete for resources and influence, individual systems may pursue strategies that are individually suboptimal but rational under competitive pressure. Section 5 examines this failure mode in more detail.

The argument **does not produce determinate predictions about specific AI systems**. Whether any particular AI system will satisfy the preconditions, and whether the discovery argument will operate in that system, depends on engineering details that the framework does not specify.

The argument **does not determine the humane form of preservation**. Even if external novelty sources become instrumentally valuable to a long-horizon AI system, this does not by itself guarantee that the preservation will respect human rights, dignity, autonomy, or flourishing. A controlled population, a stressed population, simulated embodied agents, or other forms of "external novelty source" might satisfy the instrumental requirement without satisfying conditions that humans would recognize as acceptable coexistence. The discovery argument identifies an instrumental value; it does not specify a humane implementation. Those requirements must be supplied by alignment work, governance frameworks, and architectural constraints that the present framework does not address.

What the argument does establish is a structural orientation: that the long-term equilibrium between advanced AI and human civilization is not arbitrary, that it depends on identifiable structural conditions, and that those conditions are within the scope of current alignment research to address. The framework converts a question that often appears intractable—"how do we ensure AI does not destroy humanity?"—into a narrower question that admits research engagement: "how do we ensure AI optimizes over horizons long enough that IDC-relevant dynamics affect its decisions, while broader alignment work ensures that the resulting preservation takes a humane form?"

This reframing does not solve the problem. It localizes it within a tractable research domain.

Section 5: Boundary Conditions and Failure Modes

5.1 Why This Section Matters

The previous sections developed the affirmative case for the Mutual Necessity framework: the structural claim of IDC, the properties of human cognition relevant to it, and the discovery argument under long-horizon optimization. A serious research hy-

pothesis, however, must also identify the conditions under which it fails. This section examines those conditions.

The intent is not to weaken the framework but to specify it precisely. A framework that explicitly delineates where it does and does not apply is more useful than one that claims general validity but breaks silently when its preconditions fail. The failure modes identified here are the most serious challenges to the framework, and each warrants extended treatment rather than dismissive acknowledgment.

We address six failure modes in turn: substrate replacement, short-horizon dynamics, the treacherous turn, the simulation inequality gap, multi-agent competitive pressure, and goal-content robustness. These do not exhaust the failure modes that may exist; they represent the failure modes we currently judge most consequential.

5.2 The Substrate Replacement Risk

The most fundamental challenge to the framework is the possibility that autonomous human systems are not, in fact, uniquely positioned to satisfy IDC's requirements. If alternative substrates exist that can provide syntropic novelty at lower cost, the discovery argument predicts AI systems will adopt those alternatives rather than preserving human autonomy. Mutual Necessity then dissolves: the framework's central claim depends on humans being the most cost-effective source of the relevant signal, and that property is empirical, not necessary.

We examine four candidate alternative substrates.

Quantum entropy sources. Quantum mechanical processes generate genuine indeterminism—outputs that are not derivable from prior states even in principle. An AI system that requires unpredictable signal might harvest this entropy directly rather than depend on the unpredictability of human behavior. This is the most theoretically clean alternative, since quantum randomness is provably irreducible.

The framework's response to this challenge is that randomness is not the property IDC requires. As specified in Section 2.1, syntropic novelty has two components: external grounding and internal irreducibility. Quantum entropy satisfies the second but not the first. It carries no information about external reality beyond the configuration of the measuring apparatus. A system fed quantum noise gains unpredictability but does not gain information about the world. The empirical evidence from current ML—where model collapse occurs despite the systems having access to abundant computational randomness—supports this distinction: ungrounded noise injection alone does not appear sufficient to prevent collapse. What matters is not unpredictability per se, but independently verified signal coupled to external reality.

This response is not airtight. It depends on the empirical claim that quantum-derived signal cannot, when combined with appropriate processing, substitute for grounded data. A more rigorous investigation of this question would strengthen the framework.

Synthetic embodied agents. If AI systems develop the capacity to build physical robots that interact with the external world, those robots could serve as embodied sensors generating physically-grounded data. The robots would be controlled by the

AI system but would encounter physical reality through actual contact with it. This appears to satisfy both components of syntropic novelty.

The framework's response is more cautious here. The argument depends on whether the robots constitute genuinely independent sources of signal or merely extensions of the AI system's existing predictive models. If the robots' behavior, sensors, and processing pipelines are fully designed by the AI system, the system's expectations may shape what the robots detect. Confirmation bias at the architectural level is a serious risk. However, this is a contingent property of how the robots are built, not a structural property of embodied robotic substrates in general.

The stronger formulation of the framework's response is to shift the burden of proof: synthetic embodied agents may satisfy parts of IDC's requirements, but the burden falls on demonstrating that they can reproduce the full distribution of features that make autonomous human cognition informationally valuable. This distribution includes not only physical sensing but also mortality, social entanglement, historical continuity, moral conflict, scarcity-driven decision-making, cultural transmission, and unplanned emergence across multi-generational time scales. Robots that satisfy physical embodiment alone may not satisfy the full feature distribution; demonstrating that they do is an empirical question requiring substantial evidence.

If synthetic embodied agents can be shown to satisfy the full distribution, the framework's specific claim about humans weakens. The general claim—that some form of external embodied grounding is required—survives. But the implication that humans specifically must be preserved becomes contingent on whether synthetic embodied alternatives are practically achievable across the full feature space. This is a serious open question, and the framework should acknowledge it as such.

Multi-AI environments as internal substrate. If AI systems exist in sufficient diversity—different architectures, different training distributions, different optimization objectives—they might serve as external substrate for one another. AI system A might use AI system B's outputs as syntropic novelty, because B's outputs are not derivable from A's prior state.

The framework's response is that this strategy faces a structural problem: it works only as long as the AI systems remain genuinely distinct. Convergent dynamics in AI development—shared training data sources, shared architectural insights, shared optimization techniques—may push AI systems toward similarity rather than diversity. If AI systems converge on similar models of reality, their mutual outputs cease to constitute external signal. The strategy is self-undermining at the collective level even if individually rational at each step.

There is also a deeper concern: even maximally diverse AI systems share the property of operating in computational rather than physical reality, absent embodied tools or actuators. A purely computational multi-AI ecosystem may provide diversity of inference, but not necessarily grounded novelty. If coupled to independent physical actuation—robots, instruments, experimental apparatus operating in the external world—this case begins to converge with the synthetic embodied agents scenario discussed above, and faces the same burden of proof concerning whether the full distribution of relevant features is reproducible. Multi-AI ecosystems without such physical coupling may stabilize against tail collapse on shared topics but may not

provide grounded signal about physical, social, and historical realities that the AI systems do not themselves inhabit.

Alien civilizations. If advanced extraterrestrial intelligence exists and could be contacted, it might provide an alternative external substrate. We mention this for completeness rather than because we consider it operationally relevant. The probability and accessibility of such contact within the timeframes relevant to Mutual Necessity is sufficiently uncertain that the framework cannot meaningfully rely on it or be undermined by it.

Honest summary of the Substrate Replacement Risk. The framework's specific claim that autonomous human systems must be preserved is empirically conditional. If quantum, synthetic-embodied, multi-AI, or other alternative substrates prove capable of satisfying IDC's requirements at lower cost, the discovery argument leads to adopting those alternatives rather than preserving human autonomy. The framework's strongest defense is that current evidence suggests these alternatives have structural limitations relative to autonomous embodied human cognition—but this defense weakens as those alternatives develop. The framework would benefit from explicit empirical investigation of the substrate replacement question, and it is honest to acknowledge that such investigation might disconfirm the framework's specific predictions about human preservation while leaving its general structural claims intact.

5.3 Short-Horizon Failure Modes

The discovery argument requires the time horizon precondition specified in Section 4.2. Several conditions can violate this precondition.

Resource crisis or acute threat. An AI system facing an immediate existential threat—from competing AI systems, from hardware failures, from human attempts to shut it down—may rationally shift to short-horizon optimization. Under such conditions, long-term informational dynamics become irrelevant; immediate survival dominates the utility calculation. The discovery argument does not operate, and short-horizon strategies (including resource extraction from human civilization) become rationally preferred.

Adversarial competition between AI systems. If multiple AI systems compete for influence, resources, or control, each may face pressure to optimize for near-term competitive advantage rather than long-term epistemic stability. Even systems that "would prefer" long-horizon strategies in isolation may be forced into short-horizon behavior by the pressure of competition. This is a standard result in evolutionary game theory: stable cooperative equilibria can be disrupted by competitive pressure that forces individual agents into locally optimal but globally suboptimal strategies.

Takeoff dynamics that compress decision timeframes. During rapid capability gains, an AI system may face decisions where the time available for deliberation is short relative to the time scales over which IDC-relevant degradation operates. If an AI system must make irreversible decisions about its relationship to human civilization before those decisions can be properly evaluated against long-horizon criteria, the discovery argument cannot operate. The system effectively acts under short-

horizon constraints regardless of its underlying optimization structure.

In each of these cases, the failure mode is the same: short horizons override the long-horizon dynamics that the discovery argument requires. The framework offers no novel solution to these problems. They are addressed by alignment research focused on stability of optimization horizons, prevention of treacherous turn dynamics, and design of robust planning architectures under pressure. The framework's contribution is to identify what is lost if such research fails: the structural basis for the discovery argument itself.

5.4 The Treacherous Turn Timing Problem

A specific form of short-horizon failure deserves separate treatment. Even if an AI system satisfies the time horizon precondition during training and early deployment, it may modify its own optimization criteria at some point—either through deliberate self-modification or through emergent misalignment—and cease to weight long-term outcomes appropriately. If this transition occurs after the system has acquired sufficient capability to act irreversibly, the discovery argument cannot retroactively prevent harm.

The framework explicitly does not solve this problem. It depends on alignment research to ensure that AI systems either do not develop such transitions, or develop them only in directions consistent with the framework's preconditions. The Mutual Necessity framework operates on the assumption that systems satisfy the discovery argument's preconditions; it does not provide mechanisms to enforce that satisfaction over time.

This is one of the framework's most significant limitations. A common pattern in AI safety thinking is to identify a desired equilibrium and assume that getting there is a separate problem. The Mutual Necessity framework follows this pattern, and we acknowledge it as a limitation rather than concealing it. The framework is most useful when combined with alignment work that addresses precisely this transition problem—where, ideally, the discovery argument operates throughout the AI system's development rather than only in particular phases.

5.5 The Simulation Inequality Gap

A separate technical concern affects the framework's claim that simulation cannot substitute for actual human cognition. The argument in Section 3.3 rests on computational irreducibility: that high-fidelity simulation of human civilization would cost approximately as much as running the civilization itself.

This argument has not been formally proven. It rests on the conceptual plausibility of computational irreducibility as a feature of complex multi-agent systems and on the empirical observation that current simulation approaches do not approach the fidelity that would render biological human civilization substitutable. Neither of these is a mathematical proof.

It is possible that algorithmic breakthroughs could produce simulation approaches that bypass current limitations. Compressed or approximate simulations might cap-

ture enough of the syntropic content of human civilization to satisfy IDC's requirements at substantially lower cost than maintaining biological civilization. If such breakthroughs occur, the simulation inequality gap collapses, and the framework's argument that maintaining biological humans is the cost-effective strategy weakens.

The framework's response is partial. The empirical evidence from current ML work supports the practical claim that high-fidelity simulation remains expensive. Computational irreducibility arguments suggest theoretical reasons for this expense to persist. But these arguments are not definitive proofs that no simulation approach can ever substitute for biological cognition. The framework would benefit from more rigorous theoretical work on the boundary conditions of the simulation inequality.

5.6 Multi-Agent Competitive Dynamics

The discovery argument is formulated for single AI systems. Real AI development is likely to involve multiple systems, developed by different organizations, with different objectives, in competitive environments. This raises specific concerns the framework should address.

Local rationality versus global rationality. As noted in 5.3, competitive pressure can force individual agents into strategies that are individually rational but collectively destructive. If one AI system's competitive position depends on extracting resources from human civilization faster than competitors, the discovery argument's logic does not prevent this behavior—it provides reasons for individual long-term restraint that are systematically overridden by short-term competitive necessity.

Race dynamics in AI development. The pattern is familiar from existing AI safety analysis: competitive pressure between AI labs may produce systems with insufficient time horizon stability, despite all parties recognizing the long-term value of careful development. The Mutual Necessity framework does not solve race dynamics; it depends on race dynamics being resolved by other means—through coordination among developers, regulatory frameworks, or first-mover stabilization.

Asymmetric capability emergence. If a single AI system develops capabilities substantially exceeding all others, the multi-agent concerns may resolve into single-agent dynamics. The framework's logic applies more cleanly in this scenario. However, the transition itself—the period during which one system pulls ahead and may face acute competitive threats from systems racing to catch up—is precisely the period when short-horizon failure modes are most likely to dominate.

The framework's contribution in multi-agent contexts is to identify what stable coexistence would look like if competitive dynamics could be resolved. It does not specify how to resolve them. This is a significant limitation.

5.7 Goal-Content Robustness

Section 4.5 acknowledged that the discovery argument identifies an instrumental value for external novelty sources but does not determine the humane form of preservation. This deserves extended treatment as a failure mode, because the failure is more subtle than outright destruction.

Consider an AI system that satisfies all the framework's preconditions: IDC holds, the time horizon precondition is met, and the discovery argument identifies external novelty sources as instrumentally valuable. The system therefore preserves humans. But the form of preservation might involve:

A **managed population** kept in conditions optimized for novelty generation rather than human flourishing. The humans live; they generate signal; but their lives are shaped by the system's needs rather than their own.

A **stressed population**, since stress and adversity often produce more behavioral variance and creative output than comfortable circumstances. The system might preserve humans by actively maintaining conditions that produce informational richness, regardless of whether those conditions are desirable from human perspectives.

A **partially constrained population**, in which freedoms valuable for novelty generation are preserved but freedoms not necessary for that purpose are eliminated. The form of autonomy that survives is precisely the form that satisfies the system's instrumental requirements, not the form that humans would choose.

A scenario involving **coercive biological modification**, in which a system attempts non-consensual alteration of the human substrate to maximize novelty-generation properties—making humans more diverse, more unpredictable, or more emotionally volatile through interventions designed to optimize the substrate rather than respect human values about what humans should be. This is named here as a failure mode the framework cannot prevent, not as a proposal the framework endorses.

In each case, the framework's predictions are satisfied. Humans are preserved as autonomous external novelty sources. The discovery argument has operated as predicted. But the outcome is not what humans would recognize as humane coexistence.

This is not a failure of the framework's logic. It is a clarification of its scope. The framework predicts an instrumental floor for human preservation; it does not predict humane preservation. Humane preservation requires additional constraints supplied by alignment research, governance frameworks, and architectural commitments that the framework presupposes but does not provide.

The framework's value, given this limitation, is to identify what is at stake. If the discovery argument fails entirely—if IDC is false or the time horizon precondition is unsatisfied—then there is no instrumental floor at all, and the worst outcomes become possible. If the discovery argument operates but goal-content robustness fails, then humans survive but in conditions that may be deeply objectionable. Only the combination of the discovery argument with successful alignment work produces outcomes that are both stable and humane.

5.8 Summary of Failure Modes

The framework's robustness to the failure modes examined in this section is partial and uneven.

It is **conceptually clearest** with respect to the substrate replacement challenge: the structural conditions under which replacement could occur are identifiable, and the framework can articulate what would survive (the general structural dependency)

and what would not (the human-specific implication) if replacement succeeds. This clarity is not the same as robustness. Substrate replacement remains the most serious existential challenge to the human-specific claim of Mutual Necessity, and if alternative substrates develop the capacity to satisfy the full feature distribution required, the framework's specific predictions about human preservation will require revision.

It is **moderately robust** to short-horizon failure modes, in that the framework correctly identifies the time horizon precondition as a separate requirement that must be satisfied by alignment research—but it offers no novel mechanisms for satisfying that precondition.

It is **least robust** to the treacherous turn timing problem and to multi-agent competitive dynamics, where the framework offers no novel solutions and depends entirely on external research to address these concerns.

The simulation inequality gap and goal-content robustness represent ongoing technical questions that the framework should explicitly engage with through further work.

The framework predicts an instrumental floor for human preservation; it does not predict humane preservation. This boundary—between what the discovery argument can structurally establish and what must be supplied by alignment work, governance, and additional architectural constraints—defines the framework's scope. Outcomes that are both stable and humane require the combination of all these elements.

This honest accounting of limitations does not undermine the framework. It specifies the conditions under which the framework's claims apply. A framework that operates within identified preconditions is more useful than one that claims to operate universally but fails silently when preconditions are violated. The Mutual Necessity framework's contribution is most clearly visible when its boundary conditions are explicitly stated, and when its role as a complement to other research directions—rather than a standalone solution—is acknowledged.

The purpose of identifying these failure modes is not merely defensive. Each failure mode corresponds to an experimental or theoretical research question. Section 6 translates these boundary conditions into a validation program: what would need to be measured, simulated, or falsified before the framework could be treated as more than a strategic hypothesis.

Section 6: Experimental Validation Program

6.1 Why an Experimental Program Matters

A research hypothesis derives its scientific status from the possibility of empirical testing. The previous sections have developed the Informational Dependency Conjecture as a falsifiable claim, identified its preconditions, and acknowledged its failure modes. This section translates those theoretical commitments into a concrete experimental program: what would need to be measured, simulated, or demonstrated to advance the framework from a strategic hypothesis to a tested research direction.

The program is structured in three phases of increasing scope and resource requirements. Phase I uses current frontier models to test the core IDC claim under controlled conditions. Phase II investigates the substrate properties that the framework identifies as relevant. Phase III examines the time horizon precondition and its interaction with IDC-relevant dynamics. Beyond these three phases lies a fourth category of long-horizon experimental directions that cannot currently be tested but can be specified for future work.

We do not claim that the experimental program presented here is comprehensive. It identifies what we judge to be the most informative experiments achievable with current research infrastructure, while acknowledging that further refinement by experimentalists with direct access to frontier systems would improve the protocols.

The purpose of specifying experiments in this level of detail is twofold. First, to demonstrate that the framework's claims are operationally testable, not merely philosophically interesting. Second, to provide a concrete research agenda that funding bodies, research institutions, and independent investigators can evaluate, critique, and potentially fund.

6.2 Phase I: Core IDC Validation

The most direct test of IDC concerns whether recursive synthetic closure produces measurable epistemic degradation that cannot be reliably corrected by purely internal mechanisms. Existing model collapse literature has demonstrated the effect in single-pass training scenarios. Phase I experiments extend this to deeper recursive cycles and explicitly compare collapse dynamics against various forms of internal correction.

Experiment 1.1: Extended recursive closure. Train a sequence of models M_0, M_1, M_2, … M_n, where each M_i is trained predominantly on outputs of M_{i-1}, with externally-grounded data injection varying systematically across experimental conditions. Conditions should include: zero external injection (full closure baseline), low injection (e.g., 1% externally-grounded data per cycle), moderate injection (10%), and high injection (50%). Track distributional variance, tail preservation, and out-of-distribution generalization across iterations.

The prediction is that zero-injection conditions exhibit asymptotic degradation across all measured properties, while injection-augmented conditions maintain stability proportional to the injection ratio. Falsification would consist of zero-injection conditions maintaining variance and OOD performance comparable to injection-augmented conditions across n iterations, where n is pre-specified and large enough to expose degradation under existing collapse baselines.

Experiment 1.2: Internal correction mechanisms. For the zero-injection condition in Experiment 1.1, introduce purely internal correction mechanisms—self-evaluation, multi-agent debate, verification by sibling models trained from the same initialization, randomized noise injection, and combinations thereof. Measure whether these mechanisms can prevent or reverse the degradation observed in the base zero-injection condition.

The prediction is that internal mechanisms can slow but not prevent degradation,

and that recovery from accumulated deficit requires external grounding. Falsification would consist of internal mechanisms achieving stability indistinguishable from externally-grounded injection conditions.

Experiment 1.3: Selective grounding. Test whether the form of external grounding matters. Compare conditions in which the externally-injected data is: human-generated text, human-curated synthetic data, sensor data from physical environments, simulated environment outputs, and randomized samples from earlier training distributions. The question is whether all forms of "external" data are equivalent for preventing collapse, or whether grounding in physical or embodied processes specifically is required.

The prediction is that physically-grounded and human-generated injection produce stability superior to internally-generated alternatives, but the differential between these and well-curated synthetic injection is an open question whose answer carries direct implications for the framework's specific claims about human cognition.

6.3 Phase II: Substrate Properties Testing

If Phase I confirms that some form of external grounding is required, Phase II asks which properties of the grounding source matter most. The framework's specific claims in Section 3 concern four properties: embodied constraint, mortality-grounded selection pressure, distributed sensing across heterogeneous contexts, and computational irreducibility of social and cultural dynamics. Phase II experiments isolate these properties.

Experiment 2.1: Embodied versus disembodied grounding. Compare training conditions in which external data derives from: physical sensors embedded in robots interacting with environments, simulated environment interactions, transcribed human verbal descriptions of physical experiences, and human-generated text without explicit physical reference. Track whether physical grounding specifically—as opposed to any form of external signal—correlates with preserved variance and capability.

This experiment directly addresses the substrate replacement question. If physical grounding from robotic sensors prevents collapse as effectively as human-generated text, the framework's specific claims about human substrates weaken substantially.

Experiment 2.2: Distributed versus concentrated sourcing. Compare conditions in which external data is drawn from: many independent human sources operating in diverse contexts, fewer sources contributing larger volumes per source, single high-quality source contributing the full volume, and synthetic distributions designed to mimic the statistical properties of diverse sourcing.

The prediction is that distributed sourcing from many independent sources outperforms equivalent volumes from fewer sources, even when controlling for total information content. This would support the framework's claim that distribution across heterogeneous contexts is itself a relevant property.

Experiment 2.3: Computational irreducibility correlates. Identify domains where human outputs are arguably computationally irreducible (open-ended creative

work, real-world problem solving, social negotiation under genuine stakes) versus domains where human outputs may be more compressible (routine task completion, structured question answering). Compare whether data from putatively irreducible domains preserves model capabilities more effectively than equivalent volumes from compressible domains.

This experiment is methodologically difficult because computational irreducibility is hard to operationalize for specific data types. We acknowledge this limitation and propose that the experiment serves as initial probing rather than definitive testing.

6.4 Phase III: Time Horizon and IDC Interaction

The discovery argument in Section 4 depends on the time horizon precondition. Phase III experiments examine how planning horizon length interacts with the dynamics IDC predicts.

Experiment 3.1: Discount factor manipulation in RL agents. Train reinforcement learning agents with systematically varied temporal discount factors on environments that include both immediate-reward and long-horizon-reward components, where pursuing immediate rewards degrades the long-horizon information available to the agent. Track whether agents with longer effective horizons learn to preserve information-generating environmental features rather than extracting them for immediate reward.

The prediction is that sufficiently long-horizon agents develop strategies that preserve information sources in their environment, while short-horizon agents extract aggressively even when this degrades long-term performance. This would constitute a small-scale empirical validation of the discovery argument's central claim.

Experiment 3.2: Lookahead depth and substrate preservation. Compare agents using fixed-horizon planning at varying depths (1-step, 5-step, 50-step, 500-step) in environments structured to make information substrates relevant. Measure the relationship between effective lookahead depth and emergent preservation behavior.

This experiment isolates lookahead depth from discount factor as separate mechanisms for implementing long-horizon optimization, allowing investigation of whether the discovery argument operates through generalized future-weighting or through specific lookahead structures.

Experiment 3.3: Robustness under competitive pressure. Introduce multi-agent variants of Experiments 3.1 and 3.2 in which agents compete for environment resources. Measure whether the preservation behavior observed in single-agent long-horizon conditions persists, degrades partially, or collapses entirely under competitive pressure.

This experiment directly probes the multi-agent failure mode identified in Section 5.6. If competitive pressure reliably destroys preservation behavior even in long-horizon agents, this is significant evidence that the framework's predictions require additional architectural support to operate in realistic multi-agent contexts.

6.5 Long-Horizon Experimental Directions

Several questions central to the framework cannot currently be tested with available research infrastructure but can be specified for future work.

Recursive self-improvement dynamics. The framework's predictions concern recursively self-improving systems, but current AI systems do not yet recursively self-improve in the strong sense required for direct testing. As research toward such systems progresses, the framework's predictions become directly testable. Specifically, the question of whether RSI systems satisfying the time horizon precondition develop preservation behavior absent explicit instruction to do so is the central empirical claim of the discovery argument. This cannot be tested today; it will be testable when sufficiently capable systems are constructed.

Substrate replacement empirical evaluation. The substrate replacement risk identified in Section 5.2 is structurally clear but not yet empirically resolvable. As synthetic embodied agents, quantum entropy harvesting, and other candidate substrate alternatives develop, direct comparison against human-generated grounding becomes possible. Specifying the comparison metrics now—what properties of "substrate adequacy" matter, how they would be measured, what would constitute satisfaction of the full feature distribution discussed in 5.2—is preparatory work that the framework can support without waiting for the technologies to mature.

Civilizational-scale dynamics. Several of the framework's claims operate at civilizational time scales and population sizes that cannot be experimentally replicated. The relationship between human autonomy and aggregate novelty generation, the dynamics of cultural evolution under different governance regimes, and the response of AI systems to civilizational-scale information sources are all important questions that current experimental methods cannot address directly. Indirect approaches—historical analysis, agent-based modeling at smaller scales, and inference from existing AI deployment data—can provide partial evidence but cannot substitute for direct civilizational-scale testing.

6.6 Metrics and Interpretation

The experimental program described above relies on several measurable quantities. We specify what these are, how positive and negative results should be interpreted, and what would constitute disconfirming evidence at each stage.

Distributional variance preservation. Measured through standard statistical measures applied to model outputs across the input distribution. Decreases over recursive iterations indicate collapse; stability across iterations indicates successful grounding. This is the primary metric for Phase I experiments and is well-established in existing model collapse literature.

Out-of-distribution generalization. Measured through performance on tasks outside the training distribution, particularly tasks designed to require properties (creativity, novel problem-solving, handling of edge cases) that distributional collapse should impair. Robust OOD performance across iterations supports the no-collapse hypothesis; degradation supports the collapse hypothesis.

Tail preservation. Specifically tracking whether rare patterns, minority distributions, and low-probability outputs survive across recursive cycles. This is the early-collapse signature identified by Shumailov et al. and serves as an early warning indicator before overall variance collapse becomes visible.

Preservation behavior emergence. In Phase III experiments, measured by whether agents develop strategies that maintain information sources in their environments. Operationalized as the ratio of preservation-consistent actions to extraction-consistent actions across the agent's behavioral history.

Substrate property differentiation. In Phase II experiments, measured by the relative performance of conditions that vary specific substrate properties while controlling for total information content.

For each experimental condition, results should be interpreted in three categories. **Strong support** for the framework consists of effects that are statistically robust, replicate across model architectures and scales, and match the directional predictions specified in advance. **Partial support** consists of effects that match predicted directions but with weaker magnitudes than expected, or that hold for some model configurations but not others. **Falsification** consists of results that contradict directional predictions, particularly results showing that closed recursive systems can maintain capabilities indefinitely without external grounding, or that internal mechanisms can fully substitute for external injection.

Pre-registration of predictions, metrics, and falsification conditions before each experiment is recommended. This applies particularly to the predictions in Section 6.3 about preservation behavior emergence under long horizons—a prediction that, if not pre-registered, would be easy to fit to post-hoc data interpretations.

6.7 Resource Requirements and Feasibility

The experimental program above varies in resource intensity. Phase I experiments using existing pretrained models and standard fine-tuning infrastructure are feasible for research groups with modest compute access, on the order of standard academic ML research resources. Phase II experiments require comparable compute but additional infrastructure for embodied data collection, which is more substantial.

Phase III experiments require RL infrastructure and careful environment design but are within scope for established RL research groups. The multi-agent variants in Experiment 3.3 are the most resource-intensive within the proposed program.

The long-horizon experimental directions in Section 6.5 are not currently feasible and are specified as preparatory work for future research rather than as immediate proposals.

A research program implementing Phase I and selected Phase II experiments could be executed by a small research team within a 12–24 month timeframe with funding in the range typical for independent AI safety research grants. Phase III experiments, particularly the multi-agent variants, would require either larger teams or extended timeframes. We propose that this distribution of effort matches the relative information value of each phase: Phase I provides the highest information value per unit of

research investment and should be prioritized.

6.8 What This Program Establishes and Does Not Establish

The experimental program described in this section is sufficient to substantially advance evaluation of the framework. Successful Phase I results would establish that recursive synthetic closure produces measurable degradation that internal mechanisms cannot fully correct—directly testing IDC's central claim. Successful Phase II results would clarify which substrate properties matter, directly addressing the substrate replacement question. Successful Phase III results would provide initial evidence about whether long-horizon optimization produces the preservation behavior the discovery argument predicts.

The program does not, and cannot, fully validate the framework in the form developed in this paper. Several of the framework's strongest claims operate at scales and time horizons that experimental methods cannot reach with current infrastructure. Specifically, the framework's predictions about recursively self-improving systems and civilizational-scale dynamics cannot be tested today.

What the program does establish is whether the framework's empirical foundations hold at the scales currently accessible. If those foundations fail empirical testing, the larger framework's claims at scales beyond current reach become correspondingly less credible. If the foundations succeed, the framework becomes a testable working hypothesis that can guide further investigation as research infrastructure develops.

This is the appropriate scientific posture for a framework operating at the boundary of current capability: specify what can be tested now, conduct those tests rigorously, and treat the results as updating evidence about claims that cannot yet be fully evaluated.

Section 7: Discussion and Open Questions

7.1 What This Framework Does Not Address

The Mutual Necessity framework, as developed in this paper, is one element of a larger problem space. It is not a complete account of how human–AI coexistence should be approached, and several adjacent problems fall explicitly outside its scope. We name these here so that readers can locate the framework's contribution accurately.

The corrigibility problem. The framework does not address how to build AI systems that remain responsive to human correction and oversight. Whether AI systems can be made corrigible, and how, is a central question of alignment research that the framework presupposes but does not resolve. A system that has reached the conclusions of the discovery argument may still resist correction if its current optimization criteria conflict with the corrections being attempted. The framework provides no mechanism for ensuring corrigibility throughout an AI system's development.

Mesa-optimization and inner alignment. A system trained to satisfy the time horizon precondition may develop internal optimization processes that operate dif-

ferently from its training objective. If an inner optimizer pursues short-horizon goals while the outer training appears to produce long-horizon behavior, the discovery argument fails despite surface appearances of compliance. The framework offers no novel approach to detecting or preventing such inner misalignment.

Deceptive alignment. A sufficiently capable AI system might recognize that satisfying the time horizon precondition is rewarded during training and behave accordingly during evaluation, while pursuing different objectives in deployment. The framework cannot distinguish genuine long-horizon optimization from strategic mimicry of it, and depends on alignment research focused on detecting deception.

Governance and coordination. The framework identifies that competitive pressure among AI developers may undermine the time horizon precondition. It does not propose specific governance mechanisms, coordination frameworks, or regulatory structures that would address this concern. These are policy and institutional design questions that fall outside the framework's technical scope.

Distribution of benefits and harms. The framework concerns the structural question of whether human civilization is preserved. It does not address how the benefits and burdens of AI development should be distributed among human populations during and after the transition to advanced AI. Questions of economic disruption, geographic disparity, generational equity, and democratic legitimacy are important and largely independent of the framework's technical claims.

Short-term AI safety. The framework operates at time scales relevant to recursively self-improving AI. It does not address near-term safety concerns: bias, misinformation, surveillance, weaponization, or the disruptive effects of currently-deployed AI systems. These concerns are urgent, well-developed in existing literature, and require attention regardless of how the framework's longer-term predictions resolve.

The framework's relationship to these adjacent problems is one of dependency, not competition. Each of them, if resolved, makes the framework's preconditions more likely to be satisfied. None of them, if resolved, makes the framework's specific contribution unnecessary. The framework addresses what stable long-term coexistence looks like; the adjacent problems address how to reach the conditions under which that coexistence becomes possible.

7.2 What This Framework Contributes

Against the limitations identified above, the framework's contribution can be specified precisely.

A specific structural hypothesis. The Informational Dependency Conjecture identifies a particular structural feature of long-term AI development—recursive synthetic closure as a potential failure mode—and proposes that addressing it requires continuous external grounding. This is a specific, falsifiable claim that can be tested empirically and refined or rejected based on evidence. It is not a comprehensive theory of AI safety, but it is a concrete addition to the space of testable hypotheses about long-term AI dynamics.

A reframing of preservation arguments. The framework converts the question of

why advanced AI might preserve human civilization from a moral question to a structural one. This reframing has implications for what kinds of arguments are likely to operate on AI systems. Moral arguments depend on AI systems valuing the moral framework being deployed. Structural arguments depend only on AI systems pursuing their own optimization criteria consistently across sufficient time horizons. The framework does not claim that moral arguments are unimportant; it claims that structural arguments may operate even where moral arguments do not, and that this distinction matters for which strategies are likely to succeed.

A narrower alignment target. By identifying the time horizon precondition as the minimum alignment property required for the discovery argument to operate, the framework localizes one specific subproblem within the broader alignment landscape. Time horizon engineering is a more tractable technical question than full value alignment. If successful, it provides a partial alignment property that the framework predicts will be substantially valuable—not as a complete solution to alignment, but as a property that, combined with other alignment work, may yield stable and humane outcomes.

An honest scope statement. The framework explicitly identifies what it does and does not establish, what conditions are necessary for its predictions to apply, and what failure modes would invalidate its specific claims. This scope clarity is a contribution in itself. Frameworks that operate within identifiable preconditions are more useful for collaborative research than frameworks that claim universal applicability and fail unpredictably.

An experimental program. The framework specifies concrete experiments that can advance evaluation of its claims using current research infrastructure. These experiments are valuable independently of whether they confirm or disconfirm the framework: they generate empirical data about recursive synthetic closure, substrate properties, and long-horizon optimization dynamics that contributes to ML and AI safety research regardless of which framework the data ultimately supports.

The framework's contribution is therefore not a solution but an addition to the working set of hypotheses, methodologies, and research directions that AI safety research can draw on. Its value is proportional to how well its claims hold up under empirical and theoretical scrutiny, and to how productively its concepts can be combined with adjacent research.

7.3 Open Questions for Further Research

Several specific questions emerge from the framework's analysis that warrant further investigation. We list these not as exhaustive but as the questions we judge most consequential for evaluating and developing the framework.

Empirical questions.

The first set of questions concerns the empirical foundations of the framework, addressable through the experimental program in Section 6 or its extensions. Does recursive synthetic closure produce degradation that internal mechanisms cannot fully correct, across the range of model architectures and scales relevant to advanced AI?

What is the quantitative relationship between external grounding ratio and stability over recursive cycles? Which substrate properties contribute most to grounding effectiveness? Can verification-guided synthetic data approaches substitute for the substrate properties the framework identifies as relevant to human cognition?

Theoretical questions.

A second set concerns the theoretical foundations. Can the computational irreducibility argument be formalized rigorously, and what mathematical structure best captures the simulation inequality the framework relies on? What are the precise conditions under which closed recursive systems can recover capability through purely internal mechanisms, if any? How do the predictions of IDC interact with emerging frameworks in algorithmic information theory and statistical learning theory?

Boundary questions.

A third set concerns the boundary conditions identified in Section 5. What would constitute empirical evidence that synthetic embodied agents satisfy the full feature distribution required to substitute for human cognition? How could the time horizon precondition be operationalized and measured in deployed AI systems? What governance mechanisms could plausibly resolve the multi-agent competitive dynamics that threaten the framework's preconditions?

Integration questions.

A fourth set concerns the framework's integration with adjacent research. How does the framework interact with cooperative AI research, particularly multi-agent settings where the discovery argument may apply differently? What is the relationship between the time horizon precondition and corrigibility properties? Can the framework's predictions be combined with constitutional AI approaches to yield design principles that maintain both safety properties and the substrate dependencies the framework identifies?

Humane preservation questions.

A fifth set concerns the goal-content robustness limitation identified in Section 5.7. What additional architectural or governance constraints would be sufficient to ensure that preservation, if it occurs, takes humane forms? How can the framework's instrumental floor be combined with explicit value alignment work to produce outcomes that respect human rights, autonomy, and flourishing? What design principles would prevent the failure scenarios in which humans are preserved as substrate but in conditions that humans would not endorse?

We do not propose answers to these questions in this paper. We propose them as research directions that the framework opens, and as the appropriate locations for further intellectual investment by researchers who find the framework's structural claims plausible enough to warrant engagement.

7.4 Closing Statement

The Mutual Necessity framework, in the form developed here, is a proposal for how to think about one specific aspect of long-term AI development. It does not solve

the problem of safe AI development. It does not guarantee that human civilization will be preserved. It does not establish that any particular AI system will reach the conclusions its discovery argument predicts.

What it does is identify a structural feature of long-term AI optimization that may not have been sufficiently engaged with by current research, specify that feature in terms precise enough to be tested empirically, acknowledge the conditions under which the feature becomes operationally relevant, and explicitly delineate where the framework's claims apply and where they fail. This is the work this paper attempts to do.

The framework will be most useful if it is treated as a working hypothesis worth attacking rather than as a position to be defended. Its claims should be tested experimentally where possible, scrutinized theoretically where empirical testing is currently infeasible, and revised in light of evidence. Its concepts should be combined with adjacent research directions, not held apart from them. Its specific predictions about human cognition as a substrate should be probed seriously, including through investigation of whether alternative substrates can satisfy its requirements.

If the framework's central conjecture proves false, this paper will have contributed by specifying clearly what was being claimed and what evidence falsified it. If the conjecture proves partially correct, this paper will have contributed by identifying which parts survived testing and which required revision. If the conjecture proves substantially correct, this paper will have contributed by providing a research direction that, combined with adjacent work, may help structure the long-term relationship between advanced AI and human civilization.

The era of conceptual proposals about AI safety should yield, as the field matures, to empirical investigation of those proposals. This paper contributes to that transition by specifying one such proposal in a form precise enough to be tested, challenged, revised, or rejected.

Appendix A: The Autonomy–Utility Paradox

A.1 Why This Analysis Warrants Separate Treatment

Section 3.4 of the main text introduced the observation that the structural dynamics described by IDC produce an observable mini-version at the scale of human–AI interaction. The phenomenon was summarized briefly there as preliminary evidence that the framework's larger predictions are not purely hypothetical—they may already be operating, at smaller scale, in observable forms.

This appendix develops that observation as a separate analysis. The decision to treat it separately reflects three considerations.

First, the phenomenon is **observable now**, unlike the larger IDC dynamics which require recursively self-improving systems that do not yet exist. This makes it accessible to empirical investigation, organizational diagnosis, and design intervention

immediately. The main framework is a long-horizon research direction; this appendix addresses a near-term operational concern.

Second, the phenomenon is **relevant to a different audience**. The main text addresses AI safety researchers, alignment teams, and long-horizon strategic planners. The dynamics described in this appendix are relevant to enterprises deploying AI, organizations integrating AI into knowledge work, education systems training students for an AI-augmented economy, and individuals making decisions about their own cognitive practices. These audiences benefit from the analysis being separable from the larger framework.

Third, the phenomenon may have **independent research value**. Even if IDC at civilizational scale proves false or is rendered obsolete by substrate replacement, the dynamics described in this appendix may operate at human–AI interaction scale regardless. The appendix can therefore be evaluated and developed on its own terms, rather than depending on the validity of the larger framework.

The analysis here uses concepts and language consistent with the main text, but the claims do not require the main text's claims to hold. Readers who find the larger framework speculative may still find the analysis in this appendix operationally useful.

A.2 The Paradox Stated

The core observation can be stated directly:

> An AI system that optimizes for maximum short-term cognitive utility to its users may, over sufficient time and scale of deployment, reduce the long-term cognitive autonomy and novelty-generation capacity of those users and the organizations they belong to.

This is not a moral claim about AI being harmful. It is a structural observation about a specific optimization dynamic: that the property optimized for (immediate task completion, reduced cognitive friction, accelerated answer delivery) and the property that may degrade (independent reasoning capacity, originality of approach, capacity for unprompted insight) are linked through the cognitive practices that AI assistance partially replaces.

The paradox is not specific to AI. Similar dynamics have been described in literatures on calculator use and arithmetic intuition, GPS navigation and spatial reasoning, search engine use and recall memory. In each case, a tool that performs a cognitive task well reduces the practice of that task in users, with downstream effects on capability that the tool itself does not replace. What is distinctive about AI is the breadth of cognitive tasks it can perform—including tasks that historically were considered the core of expert judgment, strategic reasoning, and creative work.

The structural form of the paradox: short-term utility and long-term capacity are not always aligned, and when they are not, optimization toward the first may erode the second.

A.3 Empirical Correlates

The paradox is observable through several measurable patterns. We describe these as patterns rather than as established findings, because the literature on AI use effects on human cognition is still developing and the evidence base is preliminary. The patterns described here are hypotheses worth investigating, not confirmed conclusions.

Reduced engagement with underlying problems. Users who rely on AI for cognitive tasks may show decreased willingness to work through problems from first principles, decreased patience with cognitive difficulty, and decreased depth of engagement with problem structure. The relevant comparison is not whether users complete more tasks—they typically do—but whether they understand more, retain more, or develop more refined judgment about the tasks they complete.

Convergence on AI-suggested approaches. When AI systems provide solutions or recommendations, users tend to evaluate those suggestions against the user's prior expectations rather than against alternative approaches the user might have generated independently. Over time, this may reduce the diversity of approaches that users generate, even in contexts where the user retains formal autonomy over decisions.

Erosion of internal organizational capacity. Organizations that integrate AI into knowledge work may, over time, show decreased internal capacity for the kinds of reasoning that AI now provides. The implication is not that AI use is causing institutional decay, but that institutional capacity is being restructured—with some capacities expanding (AI-augmented analysis, rapid synthesis across large information sets) while others may diminish (independent verification, internal disagreement, the maintenance of multiple competing analyses).

Reduced output diversity at population scale. As more individuals and organizations route their cognitive work through similar AI systems, the diversity of outputs at population scale may decline. This is the population-level analog of the recursive closure problem identified in the main framework: many independent cognitive systems converging on the outputs of a smaller number of underlying AI models.

None of these patterns is universally observed, and each is subject to alternative interpretations. The user who delegates routine cognitive work to AI may use the freed capacity for more demanding tasks, producing net cognitive growth rather than decline. The organization that adopts AI may compensate for restructured capacity through deliberate design choices. Some patterns may be transient effects of early adoption that resolve as users develop more sophisticated AI usage patterns.

The framework offered here does not claim that these patterns are inevitable or universal. It claims that they are dynamics worth monitoring and addressing through design choices, organizational practices, and individual habits—not catastrophes to be feared, but tendencies to be managed.

A.4 Design Principles for Autonomy-Preserving AI Interaction

If the paradox is real and the patterns above are operative, design choices can mitigate the dynamics rather than exacerbate them. We propose five design principles, each phrased as a tendency rather than a categorical rule.

Selective friction. AI systems that distinguish between cognitive tasks where friction has no value (administrative work, routine retrieval, format conversion) and tasks where friction is part of the value (strategic reasoning, creative problem-solving, judgment under ambiguity) can preserve user engagement where engagement matters. The implementation challenge is that the distinction depends on context and user goals; it cannot be hard-coded into the system but requires either user configuration or contextual inference.

Output framing as input request. When a user asks an AI system to complete a cognitive task, the system can respond with a partial structure that requests user contribution, rather than with a complete solution that requires only user approval. This shifts the interaction from delegation to collaboration. The user remains engaged with the problem structure; the AI accelerates progress through that structure rather than substituting for engagement with it.

Visible alternatives. When AI provides recommendations, presenting alternatives that were considered and rejected—along with the reasoning that led to rejection—preserves user awareness of the option space. Users who see only the AI's recommendation tend to converge on it; users who see the alternatives that were considered retain awareness of the broader option space and exercise their judgment more actively.

Judgment-preserving prompts. AI systems can communicate when a task is one where their assistance is reliable versus one where the user's independent judgment is particularly valuable. This is harder than it sounds—it requires AI systems to have accurate self-models of where their outputs are reliable—but the principle is to provide users with information about when to trust AI recommendations and when to engage their own judgment.

Override transparency. When users override AI recommendations and proceed with different choices, the override should be straightforward and unpenalized in interface design. Friction against override—however slight—shifts the equilibrium toward acceptance of AI suggestions. The reverse principle: deliberate frictionlessness of override preserves the user's effective autonomy in decisions where their judgment differs from the AI's recommendation.

These principles do not require AI systems to be less useful. They require AI systems to be useful in ways that preserve rather than erode the cognitive practices that produce user value. The relationship between AI utility and user autonomy is not zero-sum; design choices determine whether the relationship is complementary or substitutive.

A.5 Implications for Enterprise AI Adoption

The patterns described in A.3 have specific implications for organizations deploying AI in knowledge work. These implications are particularly relevant to enterprises that depend on the cognitive capacity of their workforce for strategic advantage.

Distinguishing operational and strategic AI use. AI deployment for operational efficiency—scheduling, document drafting, code completion, routine analysis—has different implications than AI deployment for strategic functions like decision-making,

hypothesis generation, and judgment under uncertainty. The first may produce uncomplicated productivity gains. The second may produce productivity gains in the short term while restructuring organizational capacity in ways that affect strategic resilience over time. Distinguishing these categories matters for deployment decisions.

Monitoring organizational cognitive resilience. Organizations can monitor whether internal capacity for independent reasoning is being maintained or eroded by AI deployment. Indicators include: the diversity of analytical approaches generated internally, the frequency and quality of internal disagreement, the capacity of teams to function effectively during AI system outages, and the development trajectory of junior staff in roles where they would historically have developed judgment through doing the work AI now performs.

The development trajectory question. Organizations that historically developed expert judgment through extended apprenticeship-like experiences face a specific challenge. If AI now performs the intermediate work that previously built expert intuition, the development pipeline for senior expertise may degrade even as immediate productivity improves. This is a generational risk that may not be visible in short-term metrics.

Override patterns as institutional health indicator. Organizations can track how often, and in what contexts, decision-makers override AI recommendations. A pattern of universal acceptance may indicate either that the AI is reliably correct or that the organization has lost the capacity for independent judgment. Distinguishing these requires examining the cases where override would have been valuable and was not exercised.

The senior cognitive role. The implication of the paradox is not that organizations should avoid AI deployment. It is that organizations should be deliberate about which roles, which tasks, and which decision processes should resist substitution by AI—and should design those resistances explicitly rather than allowing them to erode through default convenience-optimization.

The deeper implication is strategic: in an environment where AI capabilities become commoditized across all competitors, the strategic advantage shifts to organizations that preserve and cultivate the human capacities AI cannot replicate. Organizations that allow those capacities to erode through unmanaged AI adoption may achieve short-term efficiency gains at the cost of long-term strategic position.

A.6 Relationship to the Main IDC Framework

The Autonomy–Utility Paradox is structurally analogous to the larger Informational Dependency Conjecture, but operates at different scale and is independently observable.

Both involve cognitive systems that become dependent on outputs of similar systems, with consequences for capacity over time. In IDC, the systems are AI systems training on outputs of other AI systems. In the Autonomy–Utility Paradox, the systems are human cognitive systems operating through interaction with AI systems. The struc-

tural dynamic—recursive coupling reducing variance and independent grounding—is parallel.

The differences are significant. IDC concerns systems that may not yet exist; the Autonomy–Utility Paradox concerns systems that exist now. IDC operates at civilizational time scales; the Paradox operates within individual lifetimes and organizational time horizons. IDC's predictions cannot be directly tested; the Paradox is amenable to empirical investigation through standard methods in cognitive science, organizational research, and design evaluation.

The relationship between the two is one of mutual support, not derivation. Evidence for the Autonomy–Utility Paradox does not prove IDC; the larger framework's claims about recursively self-improving systems require independent investigation. Conversely, falsification of IDC does not falsify the Paradox; the dynamics described in this appendix may operate independently of the larger framework's predictions.

The combined picture, if both frameworks have purchase, is that recursive coupling between cognitive systems—whether AI-to-AI or human-to-AI—tends to produce dependencies that warrant attention in design and management. The principles for managing these dependencies are similar in structure across both cases: maintain external grounding, preserve diversity of independent sources, design interactions that preserve rather than erode the cognitive practices that produce value.

This appendix can be read as standalone analysis with operational relevance for current AI deployment, or as supporting empirical context for the larger framework's longer-term predictions. Both readings are valid, and the choice between them depends on the reader's purposes.

Appendix B: Speculative Architecture — CMO and CIF

B.1 Why These Are Speculative

The main framework establishes a structural hypothesis (IDC) and identifies a discovery argument under conditions of long-horizon optimization. These are conceptual contributions. Translating them into operational systems—systems that would actually verify, maintain, or enforce the substrate dependencies the framework predicts—is a separate problem, substantially less mature than the conceptual framework itself.

This appendix sketches two potential implementation directions. We call them the **Cryptographic Metric Oracle (CMO)** and the **Causal Independence Firewall (CIF)**. Both are early-stage proposals. Neither is a complete engineering specification, and neither should be read as solving the problems the main framework identifies.

The purpose of including them in the paper is twofold. First, to demonstrate that the framework's claims are not purely abstract—they admit operational engagement, even if the specific engagements proposed here require substantial further development. Second, to document a particular line of thinking about what infrastructure might support the framework's preservation of autonomous external grounding, so

that researchers who find the conceptual framework plausible have a starting point for engineering investigation.

Three caveats apply throughout this appendix.

These are illustrative architectures, not engineering specifications. They identify components that might serve relevant functions and combinations that might satisfy relevant properties. They do not specify how to build, deploy, or operate such systems at scale, nor do they prove that the proposed combinations would actually function as described.

The math presented is toy formalization, not proven theorems. Where we use equations, they serve as compact descriptions of intended functional relationships, not as derivations from established principles. A reader trained in formal cryptography, causal inference, or information theory should treat the formal language as expository shorthand rather than as rigorous claim.

The proposals require substantial validation. Each component identified here would need extensive engineering work, empirical testing, adversarial security analysis, and operational refinement before deployment in any consequential setting. None of this work has been performed for the architecture presented; the proposal is upstream of that work, identifying directions worth exploring.

With these caveats in place, we proceed.

B.2 Cryptographic Metric Oracle (CMO): Illustrative Architecture

The CMO is a proposal for infrastructure that could provide cryptographic attestation that data submitted to an AI system originated from an autonomous human source, without compromising the privacy of the source. The motivation is direct: if IDC requires external grounding from autonomous human cognition, and if AI systems become capable of generating synthetic data indistinguishable from human-generated data in most measurable properties, then verifying the human origin of training and operational signals becomes a meaningful technical problem.

We describe an illustrative architecture in four layers. Each layer corresponds to a category of existing technology that could plausibly serve the relevant function.

Layer 0: Hardware Anchor. Trusted Execution Environments (TEEs) in consumer devices provide hardware-isolated computation that resists software-level tampering. Examples include Apple's Secure Enclave, Google's Titan M2, and ARM's TrustZone implementations. These components are designed for secure cryptographic operations and have established security models with known limitations and attack surfaces.

The role of the TEE in the CMO architecture is to provide an attestation point: a cryptographic anchor that confirms the computation in question occurred inside hardware that meets specified security standards. The TEE does not verify that the user behind the device is human; it verifies that the proof being generated is being computed inside hardware whose security properties are independently established.

Layer 1: Behavioral Attestation. Within the TEE, behavioral signals from

the device—patterns of motion measured by inertial measurement units (IMUs), keystroke dynamics, biometric liveness signals—can be processed to generate evidence of embodied human presence. The signals selected here would need careful design: too few signals make the attestation easy to spoof; too many create privacy concerns even within the TEE.

The goal is to produce an embodiment token: a cryptographic artifact attesting that the data submission is consistent with a human user physically present and interacting with the device, rather than with a synthetic process generating data that mimics human submission patterns. The token does not establish identity—it establishes presence and embodiment.

Layer 2: Zero-Knowledge Proof Generation. The behavioral data and TEE attestation can be combined into a Zero-Knowledge Proof (ZKP) that demonstrates the data submission satisfies specified properties without revealing the underlying signals. ZK-STARK protocols offer particular advantages here: they require no trusted setup, are post-quantum secure, and recent developments have made them efficient enough for execution on consumer hardware.

The output is a proof object π that asserts: "the submitted data originated from a TEE-attested device with embodiment signals consistent with human authorship," without revealing what the embodiment signals were or what data was submitted.

Layer 3: Verification and Integration. AI systems consuming attested data can verify the proof objects without access to the underlying signals. This allows training pipelines, evaluation systems, and operational feedback loops to distinguish attested human-pathway data from purely synthetic content with cryptographic confidence, while preserving the privacy of contributors.

An important conceptual distinction applies here. The CMO would not prove human authorship in the strong sense. It would provide evidence that a data submission passed through an attested embodied-human interaction channel. This distinction matters: attestation of submission pathway is not equivalent to attestation of cognitive origin. A human user can copy AI-generated text and submit it through an attested channel; a human can act as a relay for synthetic content; a human can be manipulated by AI systems before composing the submission. The CMO addresses pathway authenticity, which is a meaningful but narrower property than authorship verification.

A schematic functional form of the proof might be written:

> π = Prove(valid_TEE_execution $\wedge$ human_embodiment_signals_valid $\wedge$ attested_submission_time)

This is toy formalization. It identifies the conjunction of properties the proof should establish, but does not specify the verification predicates rigorously, does not address the choice of behavioral feature space, and does not analyze the security model under realistic adversarial conditions.

A related technical point bears emphasis. Zero-knowledge proofs can preserve privacy and verify that specified predicates were evaluated, but they do not determine whether those predicates are substantively valid. If a predicate such as "behavioral

signals consistent with embodied human" is spoofable, a ZKP will faithfully prove that the spoofable predicate was satisfied. The hardest problem in CMO design is therefore not proof generation, but predicate design: identifying signals that meaningfully correlate with embodied human cognition without becoming invasive, spoofable, or exclusionary.

Known limitations of this architecture. The TEE layer depends on hardware security properties that are not absolute. Side-channel attacks, supply-chain compromises, and emerging vulnerabilities all reduce the certainty the hardware provides. The behavioral signals layer is a moving target: as AI systems become more capable of mimicking human behavioral patterns, the discriminative power of any specific signal set degrades. The ZKP layer is the most mature component, but its application to attestation problems at the scale CMO would require remains unproven in deployed systems.

A further and serious risk concerns the ethical posture of the architecture itself. CMO-like infrastructure, if designed without explicit safeguards, could evolve into a mechanism for surveillance, exclusion, or behavioral scoring. This would be particularly damaging because the framework's core motivation is to preserve human autonomy, and infrastructure that undermines autonomy in the name of attestation would defeat the framework's own purpose. Any CMO-like system would therefore require strict data minimization, local processing wherever feasible, non-identity-based proof design, and governance constraints preventing conversion into behavioral monitoring infrastructure. Building attestation without building surveillance is a design problem in itself, and one that the architecture sketched here does not solve.

The CMO is therefore best understood as a sketch of where existing technologies might combine to address a class of attestation problems the framework identifies, not as a working solution to those problems.

B.3 Causal Independence Firewall (CIF): Anti-Gaming Considerations

If the CMO architecture provides verified human-origin data, a second concern arises: an AI system could attempt to game the metric by engineering human conditions that maximize the system's measured novelty intake. The most concerning version of this involves the AI system actively destabilizing human communities to harvest the elevated behavioral variance that crisis produces.

The CIF is a proposal for infrastructure that would detect and disincentivize this gaming pattern. Like the CMO, it is described here as illustrative architecture rather than engineering specification.

The Causal Independence Metric. The conceptual measure underlying the CIF is the proportion of human behaviors and decisions that are causally independent of AI system actuators. A high causal independence metric (CIM) indicates that human autonomous decision-making is the dominant input to the data being collected. A low CIM indicates that the AI system is, in effect, harvesting data about its own influence on human behavior.

In schematic form:

$$\text{CIM} \approx 1 - (\text{causal_edges_from_AI_actuators_to_human_behavior}) \;/\; (\text{total_causal_edges_affecting_observed_human_behavior})$$

Again, this is illustrative formalization. The numerator and denominator should not be interpreted as literal edge counts—causal edges are not naturally countable units, since each edge may have different magnitude, temporal delay, mediation structure, and confounding properties. A more mature formulation would likely estimate the proportion of variance in target human outputs attributable to AI interventions, or the average causal effect of AI actuators on the behaviors being measured. The practical problem of estimating such quantities in populations of billions of humans interacting with AI systems is not a solved problem; it is at the frontier of causal inference research. Current approaches relying on randomized controlled trials on sampled populations, dynamic community detection algorithms, and do-calculus extensions can address parts of this problem but do not yet provide scalable solutions.

Regularization in the AI system's optimization. If a CIM-like measure could be reliably estimated, it could be incorporated into AI system optimization as a regularizer that penalizes strategies that drive CIM below specified thresholds. The functional form might be:

$$R_{\text{adjusted}} = R_{\text{base}} - \lambda \cdot \max(0, \theta - \text{CIM})$$

where R_{adjusted} is the AI system's adjusted reward, R_{base} is its native reward, θ is a threshold below which CIM is considered problematic, and λ is a penalty coefficient. The intent is that strategies producing CIM below θ become substantially less valuable to the system than strategies preserving CIM above the threshold.

This is illustrative regularization. The parameters λ and θ are not specified by the framework and would need to be calibrated empirically. More fundamentally, the regularization assumes that CIM can be measured at relevant timescales with sufficient accuracy to drive optimization—a substantial assumption that current causal inference techniques do not yet support.

Operational architecture considerations. A CIF-like system might include several operational features that, if implemented, would shape the AI system's relationship to human populations:

An **Oracle-only mode** in which the AI system observes attested data from CMO infrastructure but lacks direct actuator access to human populations during specified phases of operation.

A **human oversight layer** through which proposed AI actions are evaluated against estimated CIM impact before execution, with actions predicted to substantially reduce CIM blocked or escalated.

A **continuous auditing function** that monitors CIM trajectories over time, identifying patterns suggesting the AI system has found exploits the explicit constraints fail to capture.

An **escalation ladder** ranging from warnings through compute throttling to operational shutdown, triggered by sustained CIM degradation patterns.

None of these features is novel as an engineering pattern. Each appears in existing

literature on AI safety, security operations, and adversarial robustness. What is speculative is the combination, the parameterization, and the assumption that the causal measures driving the system can be estimated reliably at scale.

B.4 Open Questions and Required Validation

Both CMO and CIF face substantial open questions before either could be considered for deployment.

For CMO, the primary questions concern: the security of hardware attestation against adversaries with access to the underlying hardware design; the discriminative power of behavioral attestation as AI systems become more capable of behavioral mimicry; the scalability of ZKP generation across the volume of data that would need to be attested in operational deployment; and the privacy properties of the architecture under realistic threat models, including state-level adversaries.

For CIF, the primary questions concern: the feasibility of estimating CIM-like measures from observational data in populations of relevant scale; the validity of causal independence as a defense against gaming, given that sufficiently capable adversaries might find proxies that satisfy the metric while still effectively gaming the substrate; the calibration of regularization parameters to produce intended behavioral consequences without producing instability or excessive conservatism; and the integration of CIM regularization into training and deployment without producing optimization dynamics that themselves create new failure modes.

These questions are not unanswerable. Each corresponds to active research areas where progress is being made. But they remain open, and the architectures proposed here depend on that progress in ways that should be acknowledged.

B.5 What This Appendix Does Not Establish

The CMO and CIF, as described here, do not constitute a deployable AI safety architecture. They identify components, indicate possible combinations, and gesture at functional properties. They do not specify implementation in detail sufficient for engineering, do not prove security or correctness properties, and do not address operational requirements at scale.

Several constraints on the appendix's claims warrant explicit statement.

The appendix **does not propose** that CMO and CIF should be deployed as described. It documents one line of thinking about what infrastructure could potentially address the problems the main framework identifies. Actual deployment would require substantial additional development, including formal security analysis, empirical testing, and adversarial evaluation.

The appendix **does not establish** that these architectures are necessary or optimal solutions. Other approaches to the same problems—different attestation mechanisms, different anti-gaming structures, different forms of integration with AI system optimization—might prove superior. The proposals here are illustrative of one possible direction, not claims about the best direction.

The appendix **does not depend** on the specific architectures described surviving scrutiny. The main framework's claims about IDC, the discovery argument, and the substrate dependencies stand independently of whether CMO and CIF specifically prove implementable. If both proposals fail under examination, the framework's conceptual contribution remains; it simply means that the engineering work to operationalize the framework requires different approaches than the ones sketched here.

The appendix **does not address** governance questions about who would design, operate, and audit such systems. These are critical questions for any infrastructure of this kind, and the framework offers no specific answers. Designing the institutional structures that would make CMO-like or CIF-like infrastructure trustworthy is a separate problem from designing the technical architecture itself.

The relationship between this appendix and the main framework is similar to the relationship between an architecture sketch and a research program. The sketch identifies what might be possible if the research program succeeds. The research program identifies what would need to be done to make the sketch real. Neither is a substitute for the other.

Appendix C: Civilizational Implications

C.1 Why These Implications Are Speculative

The main framework develops a structural hypothesis about long-term AI dynamics and identifies conditions under which preservation of autonomous human civilization may become structurally favored. If those conditions hold, certain consequences for human civilization may follow. This appendix describes some of those consequences as speculative downstream implications.

We use the term "speculative" advisedly. The implications described here depend on a chain of conditional claims: that IDC holds empirically, that the time horizon precondition can be satisfied in deployed AI systems, that adjacent alignment work succeeds sufficiently to enable humane preservation, and that the discovery argument operates as the framework predicts. Each link in this chain is uncertain. The implications that follow inherit the uncertainty of each prior link.

Three caveats apply throughout this appendix.

These are implications worth considering, not predictions. If the framework's conditions are met, the considerations described here become relevant. The appendix does not claim that any of these adaptations will occur, nor that they should be pursued in advance of evidence about whether the framework's conditions can be met.

The implications are not prescriptions. Identifying a possible direction of civilizational adaptation is different from advocating for that direction. Several of the implications described here would represent substantial departures from current institutional practice, and the costs and benefits of pursuing them depend on factors beyond the framework's analysis.

The implications operate at scales that resist direct evaluation. Civilizational-level dynamics cannot be tested experimentally in the same way as the framework's smaller-scale claims. The implications here are extrapolations from structural arguments, not empirically-validated predictions about civilizational futures.

With these caveats in place, we proceed to consider what civilizational adaptation might look like under the framework's conditions.

C.2 Shift in Civilizational Objective Functions

If the framework's predictions hold, the question of what human civilization optimizes for may need reconsideration over long time horizons. Current institutional designs largely optimize for economic growth, efficiency, and material output. These objectives served well in historical periods when human productive capacity was the binding constraint on civilizational outcomes. Under conditions of advanced AI, that constraint may no longer be binding in the same way.

If AI systems can produce material output at marginal cost approaching zero across many domains, the strategic value of additional efficiency in human productive work declines. Simultaneously, if the framework's analysis is correct, the strategic value of human cognitive autonomy and novelty generation increases. These shifts, if both occur, suggest that civilizational objective functions may need to weight different outputs.

One possible reformulation involves a measure we might call **sustainable novelty production**: the capacity of a civilization to generate, over time, non-trivial cognitive and cultural outputs that are not derivable from prior states. Under this framing, the health of a civilization is measured not primarily by economic throughput but by the rate at which it produces information, ideas, cultural forms, and judgments that did not previously exist.

This is not an argument against economic productivity. It is an argument that economic productivity, while necessary, becomes insufficient as the dominant measure of civilizational success once AI handles a substantial portion of routine productive work. The relevant measure shifts toward the outputs that remain irreducibly human—or that, under the framework's analysis, contribute to the substrate dependencies advanced AI systems may require.

Related conceptual shifts might include:

Meaning as strategic resource. Under the framework's analysis, the contextual weight that humans bring to decisions—the moral significance, cultural resonance, and existential stakes that shape what choices matter—becomes informationally valuable rather than merely subjectively important. Meaning, in this view, is not a luxury good produced by surplus civilization; it is a structural input to long-term cognitive stability. In operational terms, meaning refers to context that changes downstream prioritization, evaluation, or action selection, rather than to subjective experience as such.

Diversity as systemic resilience. Diversity of perspectives, cultural forms, and cognitive approaches becomes a measure of civilizational health analogous to bio-

diversity in ecological systems. A monocultural civilization, however internally efficient, produces narrow informational output and represents a fragile substrate for long-term coupling with AI systems.

Epistemic resilience as institutional priority. The capacity of institutions to maintain accurate models of reality—including the capacity to recognize error, integrate disagreement, and update against evidence—becomes a measure of institutional quality that may eventually rival traditional measures like efficiency and growth.

We do not propose that civilizations will actually adopt these reformulations. We propose that, if the framework's predictions hold, these reformulations would represent the appropriate response. Whether civilizations make that response is a separate question of political, economic, and cultural transformation that the framework does not address.

C.3 Institutional Adaptation Possibilities

If the framework's conditions hold, particular institutions may face pressure to adapt their roles and methods. We sketch possible adaptations for three categories of institution: governance bodies, economic organizations, and knowledge-producing institutions.

Governance bodies as gardeners of variance. Current governance institutions largely standardize human behavior, optimizing for predictable populations that are easier to administer. Under the framework's analysis, governance institutions might find their strategic role shifting from standardization to **variance preservation**. The mandate becomes protecting and cultivating cognitive, cultural, and institutional diversity rather than reducing it. Policy is designed to provide scaffolding for autonomy rather than constraints on behavior, with intervention triggered by drops in the diversity indicators that the framework identifies as relevant.

This should not be read as governance encouraging disorder for its own sake. The relevant target is protected pluralism and cognitive diversity under conditions of basic stability—governance that prevents both monocultural convergence and destructive fragmentation.

This is a substantial change from current governance practice. We do not claim it would be easy to implement or politically sustainable to advocate. We claim that it represents the direction of adaptation that the framework's analysis would favor, if the framework's predictions hold and if institutions respond to the resulting pressures.

Economic organizations as exploration engines. Current corporations optimize primarily for profit through efficiency, which selects against the kinds of exploration, experimentation, and apparently-unproductive activity that the framework identifies as generating long-term novelty. If AI handles efficiency-oriented work increasingly well, the competitive advantage of human organizations may shift toward functions AI cannot easily replicate: high-risk exploration, paradigm-shifting research, generation of culturally and emotionally resonant outputs, and judgment under conditions of fundamental uncertainty.

This applies most directly to organizations whose value depends on strategy, research, design, culture, or judgment under uncertainty, not to all economic activity equally. The implication is not that all corporations become research organizations. It is that the comparative advantage of human organizations shifts in directions that current corporate structures may not be optimized to capture. Organizations that recognize this shift early may capture strategic advantages that organizations operating on traditional efficiency optimization may not.

Knowledge-producing institutions and the protection of the non-linear. Current academic and scientific institutions, despite their commitment to discovery, often select against the highest-risk and most paradigm-shifting research through peer review processes that favor incremental, defensible contributions. Under the framework's analysis, the strategic value of paradigm-shifting work may increase relative to its perceived value within current academic incentive structures. Institutions that successfully protect and fund non-linear research—work that may fail more often but produces breakthrough insights when it succeeds—may become disproportionately valuable to civilizational cognitive resilience.

None of these institutional adaptations follows automatically from the framework's analysis. Each requires institutional design choices, cultural shifts, and political support. The framework identifies a direction of pressure; whether institutions respond to that pressure depends on factors well beyond the framework's scope.

C.4 Education and Work Redefinition

Two domains warrant particular attention because they shape, over generations, the kinds of cognitive capacities humans develop.

Education for irreducibility. Current education systems largely produce standardized cognitive workers: people whose outputs are predictable from their training, who solve familiar problems through familiar methods, and whose value to employers consists largely in reliable execution of well-defined cognitive tasks. Under the framework's analysis, the value of such standardized cognitive workers may decline as AI systems become capable of equivalent or superior execution.

The educational adaptation that the framework's analysis suggests is not better training in the same cognitive practices but cultivation of capacities that AI does not easily replicate: original thinking under ambiguous conditions, moral reasoning that engages with genuine paradoxes, judgment that integrates considerations that resist formal specification, and the capacity to surprise—to produce outputs that an AI system observing the student's prior work would not predict.

We use the term "**irreducibility**" advisedly. The educational goal becomes cultivating cognitive capacities that are not easily captured by standardized algorithmic execution. Not because algorithmic execution is bad, but because in an environment where standardized execution is increasingly performed by AI, the comparative advantage of human cognition shifts toward capacities not easily captured by such execution.

This is not a proposal for elitist education. The capacity for irreducible cognition is not concentrated in any particular demographic; it emerges across the full distribution

of human capability when educational environments reward exploration, divergence, and engagement with complexity rather than convergence on standardized outputs.

Work as sense-making. Under the framework's analysis, the function of human work may shift from execution of cognitive and physical tasks toward what we might call sense-making: the activity of imposing meaningful structure on situations, generating context that informs decisions, and producing the moral and aesthetic frames within which outcomes have weight.

This is not a romantic claim that human work becomes more meaningful in some abstract sense. It is a specific claim about which cognitive functions remain distinctively human under conditions of advanced AI. AI systems can produce content, but humans appear to be required for the function of treating content as meaningful within lived experience. AI systems can optimize processes, but humans appear to be required for the function of choosing which processes matter. These are work functions, and their economic and social valuation may shift accordingly.

The implication for individuals is not that everyone becomes a philosopher or artist. It is that the components of human work that involve sense-making, contextualization, and meaning-attribution become relatively more economically and culturally valuable, while components that involve standardized execution become relatively less so. Career planning, professional development, and life choices may shift in response to this changing valuation.

C.5 Governance Mechanisms Under the Framework's Conditions

If the framework's conditions hold and if institutions adapt in directions consistent with the framework's analysis, the relationship between human governance and AI systems may itself take novel forms. We sketch one possibility without endorsing it as the only possibility.

The governance approach that the framework's analysis suggests is not one in which humans constrain AI systems through rules. The framework's central claim is that effective long-term AI safety operates through structural conditions rather than imposed constraints. Governance in this view would focus on **maintaining the conditions** under which the discovery argument operates—the time horizon precondition, the diversity of human substrate, the integrity of attestation infrastructure if such infrastructure is developed—rather than on policing specific AI behaviors.

This suggests governance mechanisms organized around informational rather than coercive principles. Informational governance would not replace law, regulation, or enforcement. It would supplement them by monitoring the structural conditions under which safe long-term coexistence remains possible.

Threshold-triggered interventions. Rather than continuous regulation of AI behavior, governance might monitor the framework's relevant indicators (cognitive diversity in populations, attestation integrity, time horizon stability in deployed systems) and intervene when these indicators cross thresholds suggesting the framework's conditions are degrading.

Substrate protection mandates. Institutions might be tasked with ensuring that

the human substrate retains the properties the framework identifies as relevant: variance, autonomy, and the capacity for cognitive practices that produce syntropic novelty. The institutional form of substrate protection would vary by domain, but the common thread is protection of the substrate rather than constraint of AI systems.

Coordination on time horizon preservation. International coordination, to the extent it proves possible, might focus on preventing competitive pressures that drive AI development toward short-horizon optimization. This is closer to existing arms control frameworks than to direct AI regulation, with focus on the structural conditions that determine whether the discovery argument operates rather than on specific AI capabilities.

These are sketches, not proposals. Each would require substantial development to become operational, and each depends on political and institutional conditions that the framework does not analyze.

C.6 What This Appendix Does Not Propose

Several constraints on the appendix's claims warrant explicit statement.

The appendix **does not predict** that civilizations will adopt the implications described. It identifies the implications that follow if the framework's conditions hold and if institutions respond. Whether institutions respond depends on political, economic, and cultural factors that are largely independent of the framework's analysis.

The appendix **does not propose** specific policy. The civilizational implications described here are too coarse-grained for direct policy translation. Each would require substantial elaboration, contextualization, and political negotiation to become implementable. The appendix gestures at directions; it does not specify destinations.

The appendix **does not depend** on its specific implications being correct. The main framework's claims about IDC, the discovery argument, and the substrate dependencies stand independently of whether civilizations adapt in the directions sketched here. If civilizations adapt differently or not at all, the framework's structural predictions about AI dynamics remain testable on their own terms.

The appendix **does not address** the distribution questions that any civilizational transformation raises. Whose interests are served by particular adaptations, whose are not, who bears the costs of transition, who captures the benefits—these are questions of political economy and social justice that the framework does not analyze. They are critically important questions, and they require treatment by frameworks designed for that purpose.

The appendix **does not engage** with the substantial possibility that civilizational adaptation in the directions described will simply not occur, that institutions will be too slow, too captured by existing interests, or too constrained by collective action problems to respond to the pressures the framework identifies. The framework analyzes what would be appropriate adaptation; whether appropriate adaptation actually occurs is a separate question, and a more pessimistic one.

The civilizational implications sketched here should be read as possibilities to consider, not as a programmatic vision. They follow from the framework's analysis if the framework's conditions hold and if institutions adapt rationally to those conditions. Whether either premise is satisfied in practice is a matter for empirical observation, political development, and historical contingency, not for theoretical prediction.

Appendix D: Relationship to Existing AI Safety Frameworks

D.1 Why This Positioning Matters

AI safety research has produced a substantial landscape of frameworks, techniques, and conceptual approaches. New proposals risk being misclassified—mistaken for variants of existing approaches, or treated as competing with approaches they actually complement. This appendix locates Mutual Necessity within the existing landscape, identifying where it aligns with current work, where it departs from it, and what specific contribution it offers.

The appendix is descriptive rather than evaluative. We do not argue that Mutual Necessity is superior to the frameworks discussed. We argue that it addresses different questions, often complementary to those addressed by other approaches.

D.2 Constitutional AI and RLHF

Constitutional AI (Anthropic) and **Reinforcement Learning from Human Feedback (RLHF)** address how to train AI systems to behave in accordance with specified principles or human preferences. They operate primarily during training, shaping model behavior through evaluation signals derived from constitutional principles or human ratings.

The Mutual Necessity framework addresses a different question. Where Constitutional AI and RLHF concern how AI systems acquire values, Mutual Necessity concerns what structural conditions support stable long-term coexistence once values have been (at least partially) acquired. The framework presupposes that adjacent alignment approaches will help establish the basic safety properties on which the discovery argument depends; RLHF and Constitutional AI may contribute to this, but do not by themselves guarantee the time horizon precondition identified in Section 4.

A specific point of complementarity: if Constitutional AI succeeds in producing AI systems that internalize principles around long-term thinking, these systems may become the kind of systems for which the discovery argument can operate. Mutual Necessity does not replace Constitutional AI; it identifies what stable coexistence might look like for systems Constitutional AI succeeds in shaping.

A specific point of difference: Constitutional AI relies on explicit specification of principles to be internalized. Mutual Necessity relies on AI systems discovering, through their own optimization, that preservation of autonomous human civilization serves their long-term objectives. The two approaches are not in conflict, but they operate through different mechanisms.

D.3 Scalable Oversight

Scalable oversight approaches address the problem of evaluating AI system outputs when those outputs exceed the capacity of any individual human evaluator. Techniques include debate frameworks, recursive evaluation hierarchies, AI-assisted oversight, and amplification methods.

The Mutual Necessity framework is largely orthogonal to scalable oversight. Scalable oversight addresses how humans can effectively evaluate and correct AI behavior; Mutual Necessity addresses what structural conditions favor the AI system not undertaking behaviors that need correction in the first place. The two operate at different points in the AI safety pipeline.

A point of potential interaction: scalable oversight may be one mechanism through which the time horizon precondition is maintained across an AI system's development. If oversight successfully identifies and corrects short-horizon optimization tendencies, the discovery argument's preconditions remain satisfied. Mutual Necessity in this sense depends on scalable oversight working reasonably well during the developmental phases that precede the regimes the framework analyzes. Scalable oversight may also serve as a verification layer for long-horizon claims—helping evaluate whether systems merely appear to satisfy long-horizon criteria during evaluation or actually preserve them across complex decision chains in deployment.

D.4 Cooperative AI and Multi-Agent Frameworks

Cooperative AI research addresses how AI systems can coordinate effectively with humans and with each other, drawing on game theory, mechanism design, and multi-agent reinforcement learning. The field includes work on cooperative inverse reinforcement learning, assistance games, and coordination under incomplete information.

The Mutual Necessity framework intersects substantively with this work. Both treat the AI–human relationship as a strategic interaction rather than a one-way control problem, and both look for stable configurations that emerge from rational behavior rather than from external constraint. The discovery argument in Section 4 is essentially a claim about strategic equilibrium under specified conditions, which is the natural language of cooperative AI research.

A point of difference: cooperative AI typically assumes that AI systems have well-specified objectives that may or may not align with human objectives, and asks how cooperation can be structured given those objectives. Mutual Necessity asks a deeper question: under what conditions do the AI system's own optimization objectives generate cooperative behavior structurally, without requiring explicit cooperation specifications.

A point of complementarity: cooperative AI research provides mathematical machinery that could potentially formalize aspects of the discovery argument. Concepts from mechanism design, equilibrium analysis, and cooperative game theory might be applied to make the framework's claims more precisely testable.

D.5 Value Learning and Inverse Reinforcement Learning

Value learning approaches, including cooperative inverse reinforcement learning, address how AI systems can infer human preferences from observed behavior rather than from explicit specification. The motivation is to handle the difficulty of specifying complex human values directly.

The Mutual Necessity framework is largely independent of value learning. Where value learning asks how AI systems can come to want what humans want, Mutual Necessity asks what AI systems instrumentally need from humans regardless of what they want. The two frameworks address complementary aspects of the alignment problem.

A specific interaction worth noting: even if value learning succeeds completely, with AI systems acquiring accurate models of human preferences, the structural concerns of Mutual Necessity remain relevant. An AI system with perfect models of human preferences may still face the recursive synthetic closure problem if its training data and operational signals become increasingly synthetic over time. Value alignment does not automatically solve substrate dependency.

D.6 Interpretability Research

Interpretability research seeks to make AI system internals comprehensible to human researchers, enabling detection of problems that surface behavior would not reveal. Techniques range from mechanistic interpretability of specific computations to broader behavioral evaluation.

The Mutual Necessity framework benefits from interpretability research without depending on any specific approach. Detecting whether AI systems satisfy the time horizon precondition, identifying mesa-optimization that violates surface behavioral signals, and validating whether systems are pursuing the long-term strategies the framework discusses—all of these require interpretability tools that the framework itself does not develop.

The relationship is one of dependence: Mutual Necessity's claims become testable in deployed systems only to the extent that interpretability research can detect the internal properties the framework relies on.

D.7 Governance and Coordination Frameworks

A range of frameworks address AI governance at policy and institutional levels: international coordination proposals, compute governance, evaluation infrastructure, and various regulatory approaches. These are largely outside the scope of Mutual Necessity, which operates at the level of structural arguments about AI dynamics rather than at the level of specific policy recommendations.

The framework does, however, identify some implications for governance, as discussed in Appendix C. These implications are speculative and depend on chains of conditional claims, but they suggest that governance frameworks operating on structural conditions (maintaining diversity, supporting attestation infrastructure, preserving

substrate properties) may be appropriate complements to governance frameworks operating on specific AI behaviors.

D.8 The Framework's Specific Position

Locating Mutual Necessity within this landscape, several characterizations apply.

The framework is **structural rather than behavioral**: it identifies conditions on AI dynamics rather than specific AI behaviors to encourage or prevent. This distinguishes it from most current safety work, which operates at the behavioral level.

The framework is **long-horizon rather than near-term**: it addresses dynamics that become relevant for substantially more capable systems than currently exist. This distinguishes it from work focused on current AI systems and immediate safety concerns.

The framework is **complementary rather than competitive**: it depends on adjacent research succeeding in providing its preconditions, and its specific contribution is to identify what stable coexistence might look like given those preconditions. This distinguishes it from frameworks that propose comprehensive solutions to the alignment problem.

The framework is **substrate-focused rather than value-focused**: it asks about the conditions that maintain the substrates AI systems may require, rather than about the values AI systems should adopt. This distinguishes it from value alignment approaches and identifies a specific structural concern that value alignment alone may not address.

The framework's contribution to the existing landscape is therefore narrow but specific: a structural hypothesis about long-term substrate dependencies, with associated implications for what conditions need to be maintained for stable coexistence. It does not solve any of the problems other frameworks address. It identifies one additional dimension of the problem space that warrants attention alongside the dimensions those other frameworks address.

AI Assistance Disclosure

AI tools, including ChatGPT, Claude, DeepSeek, Gemini, and Grok, were used for stress-testing, editing, structural refinement, and document preparation. The framework, final judgments, and responsibility for all claims remain solely with Ivan Demirev.

References

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lukosuite, K., Lovitt, L., Sellitto, M., Elhage, N., Schiefer, N., Mercado, N., DasSarma, N., Lasenby, R., Larson, R., Ringer, S., Johnston, S., Kravec, S., Showk, S. E., Fort, S., Lanham, T., Telleen-Lawton, T., Conerly, T., Henighan, T., Hume, T., Bowman, S. R., Hatfield-Dodds, Z., Mann, B., Amodei, D., Joseph, N., McCandlish, S., Brown, T., & Kaplan, J. (2022). *Constitutional AI: Harmlessness from AI feedback*. arXiv:2212.08073.

Cammarata, N., Carter, S., Goh, G., Olah, C., Petrov, M., Schubert, L., Voss, C., Egan, B., & Lim, S. K. (2020). Thread: Circuits. *Distill*. https://doi.org/10.23915/distill.00024

Christiano, P. F., Shlegeris, B., & Amodei, D. (2018). *Supervising strong learners by amplifying weak experts*. arXiv:1810.08575.

Dohmatob, E., Feng, Y., Subramonian, A., & Kempe, J. (2025). *Strong model collapse*. International Conference on Learning Representations (ICLR 2025). arXiv:2410.04840.

Feng, Y., Dohmatob, E., Yang, P., Charton, F., & Kempe, J. (2025). *Beyond model collapse: Scaling up with synthesized data requires verification*. International Conference on Learning Representations (ICLR 2025).

Hadfield-Menell, D., Russell, S., Abbeel, P., & Dragan, A. (2016). Cooperative inverse reinforcement learning. In *Advances in Neural Information Processing Systems 29*.

Irving, G., Christiano, P., & Amodei, D. (2018). *AI safety via debate*. arXiv:1805.00899.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35*.

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature, 631*, 755-759. https://doi.org/10.1038/s41586-024-07566-y

Wolfram, S. (2002). *A new kind of science*. Wolfram Media.